Regularizers for Structured Sparsity

Charles A. Micchelli$^{(1),(2)}$

Jean M. Morales$^{(3)}$

Massimiliano Pontil$^{(3)}$

(1) Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon Tong
Hong Kong

(2) Department of Mathematics and Statistics
State University of New York
The University at Albany
1400 Washington Avenue
Albany, NY, 12222, USA

(3) Department of Computer Science
University College London
Gower Street, London WC1E
England, UK
E-mail: {*m.pontil,j.morales*}*@cs.ucl.ac.uk*

Abstract

We study the problem of learning a sparse linear regression vector under additional conditions on the structure of its sparsity pattern. This problem is relevant in machine learning, statistics and signal processing. It is well known that a linear regression can benefit from knowledge that the underlying regression vector is sparse. The combinatorial problem of selecting the nonzero components of this vector can be "relaxed" by regularizing the squared error with a convex penalty function like the ℓ_1 norm. However, in many applications, additional conditions on the structure of the regression vector and its sparsity pattern are available. Incorporating this information into the learning method may lead to a significant decrease of the estimation error.

In this paper, we present a family of convex penalty functions, which encode prior knowledge on the structure of the vector formed by the absolute values of the regression coefficients. This family subsumes the ℓ_1 norm and is flexible enough to include different models of sparsity patterns, which are of practical and theoretical importance. We establish the basic properties of these penalty functions and discuss some examples where they can be computed explicitly. Moreover, we present a convergent optimization algorithm for solving regularized least squares with these penalty functions. Numerical simulations highlight the benefit of structured sparsity and the advantage offered by our approach over the Lasso method and other related methods.

1 Introduction

The problem of sparse estimation is becoming increasing important in statistics, machine learning and signal processing. In its simplest form, this problem consists in estimating a regression vector $\beta^* \in \mathbb{R}^n$ from a set of linear measurements $y \in \mathbb{R}^m$, obtained from the model

$$y = X\beta^* + \xi \tag{1.1}$$

where X is an $m \times n$ matrix, which may be fixed or randomly chosen and $\xi \in \mathbb{R}^m$ is a vector which results from the presence of noise.

An important rational for sparse estimation comes from the observation that in many practical applications the number of parameters n is much larger than the data size m, but the vector β^* is known to be sparse, that is, most of its components are equal to zero. Under this sparsity assumption and certain conditions on the data matrix X, it has been shown that regularization with the ℓ_1 norm, commonly referred to as the Lasso method [27], provides an effective means to estimate the underlying regression vector, see for example [5, 7, 18, 28] and references therein. Moreover, this method can reliably select the sparsity pattern of β^* [18], hence providing a valuable tool for feature selection.

In this paper, we are interested in sparse estimation under additional conditions on the sparsity pattern of the vector β^*. In other words, not only do we expect this vector to be sparse but also that it is *structured sparse*, namely certain configurations of its nonzero components are to be preferred to others. This problem arises is several applications, ranging from functional magnetic resonance imaging [9, 29], to scene recognition in vision [10], to multi-task learning [1, 15, 23] and to bioinformatics [26], see [14] for a discussion.

The prior knowledge that we consider in this paper is that the vector $|\beta^*|$, whose components are the absolute value of the corresponding components of β^*, should belong to some prescribed convex subset Λ of the positive orthant. For certain choices of Λ this implies a constraint on the sparsity pattern as well. For example, the set Λ may include vectors with some desired monotonicity constraints, or other constraints on the "shape" of the regression vector. Unfortunately, the constraint that $|\beta^*| \in \Lambda$ is nonconvex and its implementation is computational challenging. To overcome this difficulty, we propose a family of penalty functions, which are based on an extension of the ℓ_1 norm used by the Lasso method and involves the solution of a smooth convex optimization problem. These penalty functions favor regression vectors β such that $|\beta| \in \Lambda$, thereby incorporating the structured sparsity constraints.

Precisely, we propose to estimate β^* as a solution of the convex optimization problem

$$\min\left\{\|X\beta - y\|_2^2 + 2\rho\Omega(\beta|\Lambda) : \beta \in \mathbb{R}^n\right\} \tag{1.2}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, ρ is a positive parameter and the penalty function takes the form

$$\Omega(\beta|\Lambda) = \inf\left\{\frac{1}{2}\sum_{i\in\mathbb{N}_n}\left(\frac{\beta_i^2}{\lambda_i} + \lambda_i\right) : \lambda \in \Lambda\right\}.$$

As we shall see, a key property of the penalty function is that it exceeds the ℓ_1 norm of β when $|\beta| \notin \Lambda$, and it coincides with the ℓ_1 norm otherwise. This observation suggests a heuristic interpretation of the method (1.2): among all vectors β which have a fixed value of the ℓ_1 norm, the penalty function Ω will encourage those for which $|\beta| \in \Lambda$. Moreover, when $|\beta| \in \Lambda$ the function Ω reduces to the ℓ_1 norm and, so, the solution of problem (1.2) is expected to be sparse. The penalty function therefore will encourage certain desired sparsity patterns. Indeed, the sparsity pattern of β is contained in that of the auxiliary vector λ at the optimum and, so, if the set Λ allows only for certain sparsity patterns of λ, the same property will be "transferred" to the regression vector β.

There has been some recent research interest on structured sparsity, see [11, 13, 14, 19, 22, 30, 31] and references therein. Closest to our approach are penalty methods built around the idea of mixed ℓ_1-ℓ_2 norms. In particular, the group Lasso method [31] assumes that the components of the underlying regression vector β^* can be partitioned into prescribed groups, such that the restriction of β^* to a group is equal to zero for most of the groups. This idea has been extended in [14, 32] by considering the possibility that the groups overlap according to certain hierarchical or spatially related structures. Although these methods have proved valuable in applications, they have the limitation that they can only handle more restrictive classes of sparsity, for example patterns forming only a single connected region. Our point of view is different from theirs and provides a means to designing more flexible penalty functions which maintain convexity while modeling richer model structures. For example, we will demonstrate that our family of penalty functions can model sparsity patterns forming multiple connected regions of coefficients.

The paper is organized in the following manner. In Section 2 we establish some important properties of the penalty function. In Section 3 we address the case in which the set Λ is a box. In Section 4 we derive the form of the penalty function corresponding to the wedge with decreasing coordinates and in Section 5 we extends this analysis to the case in which the constraint set Λ is constructed from a directed graph. In Section 6 we discuss useful duality relations and in Section 7 we address the issue of solving the problem (1.2) numerically by means of an alternating minimization algorithm. Finally, in Section 8 we provide numerical simulations with this method, showing the advantage offered by our approach.

A preliminary version of this paper appeared in the proceedings of the Twenty-Fourth Annual Conference on Neural Information Processing Systems (NIPS 2010) [21]. The new version contains Propositions 2.1, 2.3 and 2.4, the description of the graph penalty in Section 5, Section 6, a complete proof of Theorem 7.1 and an experimental comparison with the method of [11].

2 Penalty function

In this section, we provide some general comments on the penalty function which we study in this paper.

We first review our notation. We denote with $\mathbb{R}_+$ and $\mathbb{R}_{++}$ the nonnegative and positive real line, respectively. For every $\beta \in \mathbb{R}^n$ we define $|\beta| \in \mathbb{R}^n_+$ to be the vector formed by the absolute values of the components of β, that is, $|\beta| = (|\beta_i| : i \in \mathbb{N}_n)$, where $\mathbb{N}_n$ is the set of positive integers up to and including n. Finally, we define the ℓ_1 norm of vector β as $\|\beta\|_1 = \sum_{i \in \mathbb{N}_n} |\beta_i|$ and the ℓ_2 norm as $\|\beta\|_2 = \sqrt{\sum_{i \in \mathbb{N}_n} \beta_i^2}$.

Given an $m \times n$ input data matrix X and an output vector $y \in \mathbb{R}^m$, obtained from the linear regression model $y = X\beta^* + \xi$ discussed earlier, we consider the convex optimization problem

$$\inf \left\{ \|X\beta - y\|_2^2 + 2\rho\, \Gamma(\beta, \lambda) : \beta \in \mathbb{R}^n, \lambda \in \Lambda \right\} \tag{2.1}$$

where ρ is a positive parameter, Λ is a prescribed convex subset of the positive orthant $\mathbb{R}^n_{++}$ and the function $\Gamma : \mathbb{R}^n \times \mathbb{R}^n_{++} \to \mathbb{R}$ is given by the formula

$$\Gamma(\beta, \lambda) = \frac{1}{2} \sum_{i \in \mathbb{N}_n} \left(\frac{\beta_i^2}{\lambda_i} + \lambda_i \right).$$

Note that in (2.1), for a fixed $\beta \in \mathbb{R}^n$, the infimum over $\lambda = (\lambda_i : i \in \mathbb{N}_n)$ in general is not attained, however, for a fixed $\lambda \in \Lambda$, the infimum over β is always attained.

Since the auxiliary vector λ appears only in the second term of the objective function of problem (2.1), and our goal is to estimate β^*, we may also directly consider the regularization problem

$$\min \left\{ \|X\beta - y\|_2^2 + 2\rho\, \Omega(\beta|\Lambda) : \beta \in \mathbb{R}^n \right\}, \tag{2.2}$$

where the penalty function takes the form

$$\Omega(\beta|\Lambda) = \inf \left\{ \Gamma(\beta, \lambda) : \lambda \in \Lambda \right\}. \tag{2.3}$$

Note that Γ is convex on its domain because each of its summands are likewise convex functions. Hence, when the set Λ is convex it follows that $\Omega(\cdot|\Lambda)$ is a convex function and (2.2) is a convex optimization problem.

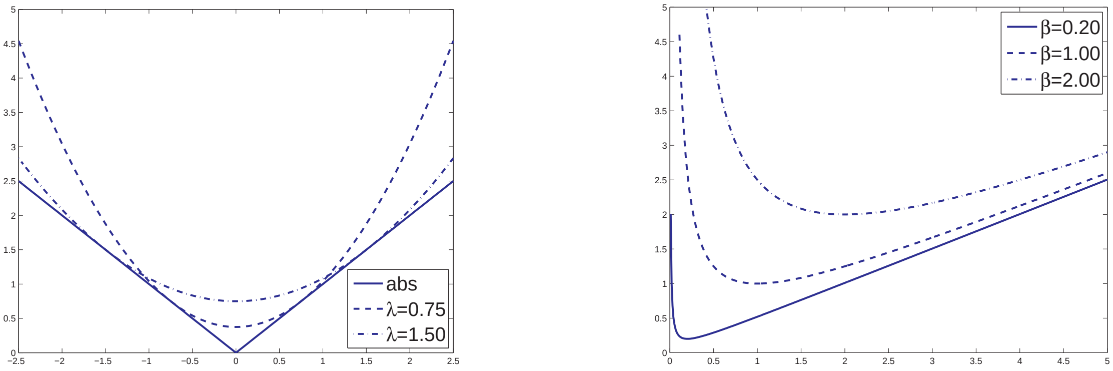
An essential idea behind our construction of the penalty function is that, for every $\lambda \in \mathbb{R}_{++}$, the quadratic function $\Gamma(\cdot, \lambda)$ provides a smooth approximation to $|\beta|$ from above, which is exact at $\beta = \pm\lambda$. We indicate this graphically in Figure 1-a. This fact follows immediately by the arithmetic-geometric mean inequality, which states, for every $a, b \geq 0$ that $(a+b)/2 \geq \sqrt{ab}$.

A special case of the formulation (2.2) with $\Lambda = \mathbb{R}^n_{++}$ is the Lasso method [27], which is defined to be a solution of the optimization problem

$$\min \left\{ \|y - X\beta\|_2^2 + 2\rho \|\beta\|_1 : \beta \in \mathbb{R}^n \right\}.$$

Indeed, using again the arithmetic-geometric mean inequality it follows that $\Omega(\beta|\mathbb{R}^n_{++}) = \|\beta\|_1$. Moreover, if for every $i \in \mathbb{N}_n$ $\beta_i \neq 0$, then the infimum is attained for $\lambda = |\beta|$. This important special case motivated us to consider the general method described above. The utility of (2.3) is that upon inserting it into (2.2) there results an optimization problem over λ and β with a continuously differentiable objective function. Hence, we have succeeded in expressing a nondifferentiable convex objective function by one which is continuously differentiable on its domain.

Our first observation concerns the differentiability of Ω. In this regard, we provide a sufficient condition which ensures this property of Ω, which, although seemingly cumbersome covers important special cases. To present our result, for any real numbers $a < b$, we define the parallelepiped $[a, b]^n = \{x : x = (x_i : i \in \mathbb{N}_n), a \leq x_i \leq b,\ i \in \mathbb{N}_n\}$.



(a) (b)

Figure 1: (a): Function $\Gamma(\cdot, \lambda)$ for some values of $\lambda > 0$; (b): Function $\Gamma(\beta, \cdot)$ for some values of $\beta \in \mathbb{R}$.

Definition 2.1. *We say that the set Λ is admissible if it is convex and, for all $a, b \in \mathbb{R}$ with $0 < a < b$, the set $\Lambda_{a,b} := [a, b]^n \cap \Lambda$ is a nonempty, compact subset of the interior of Λ.*

Proposition 2.1. *If $\beta \in (\mathbb{R}\backslash\{0\})^n$ and Λ is an admissible subset of $\mathbb{R}^n_{++}$, then the infimum above is uniquely achieved at a point $\lambda(\beta) \in \Lambda$ and the mapping $\beta \mapsto \lambda(\beta)$ is continuous. Moreover, the function $\Omega(\cdot|\Lambda)$ is continuously differentiable and its partial derivatives are given, for any $i \in \mathbb{N}_n$, by the formula*

$$\frac{\partial \Omega(\beta|\Lambda)}{\partial \beta_i} = \frac{\beta_i}{\lambda_i(\beta)}. \tag{2.4}$$

We postpone the proof of this proposition to the appendix. We note that, since $\Omega(\cdot|\Lambda)$ is continuous, we may compute it at a vector β, some of whose components are zero, as a limiting process. Moreover, at such a vector the function $\Omega(\cdot|\Lambda)$ is in general not differentiable, for example consider the case $\Omega(\beta|\mathbb{R}^n_{++}) = \|\beta\|_1$.

The next proposition provides a justification of the penalty function as a means to incorporate structured sparsity and establish circumstances for which the penalty function is a norm. To state our result, we denote by $\overline{\Lambda}$ the closure of the set Λ.

Proposition 2.2. *For every $\beta \in \mathbb{R}^n$, we have that $\|\beta\|_1 \leq \Omega(\beta|\Lambda)$ and the equality holds* if and only if $|\beta| := (|\beta_i| : i \in \mathbb{N}_n) \in \overline{\Lambda}$. *Moreover, if Λ is a nonempty convex cone then the function $\Omega(\cdot|\Lambda)$ is a norm and we have that $\Omega(\beta|\Lambda) \leq \omega\|\beta\|_1$, where $\omega := \max\{\Omega(e_k|\Lambda) : k \in \mathbb{N}_n\}$ and $\{e_k : k \in \mathbb{N}_n\}$ is the canonical basis of $\mathbb{R}^n$.*

Proof. By the arithmetic-geometric mean inequality we have that $\|\beta\|_1 \leq \Gamma(\beta, \lambda)$, proving the first assertion. If $|\beta| \in \overline{\Lambda}$, there exists a sequence $\{\lambda^k : k \in \mathbb{N}\}$ in Λ, such that $\lim_{k\to\infty} \lambda^k = |\beta|$. Since $\Omega(\beta|\Lambda) \leq \Gamma(\beta, \lambda^k)$ it readily follows that $\Omega(\beta|\Lambda) \leq \|\beta\|_1$. Conversely, if $|\beta| \in \overline{\Lambda}$, then there is a sequence $\{\lambda^k : k \in \mathbb{N}\}$ in Λ, such that $\Gamma(\beta, \lambda^k) \leq \|\beta_1\| + 1/k$. This inequality implies that some subsequence of this sequence converges to a $\overline{\lambda} \in \overline{\Lambda}$. Using arithmetic-geometric mean inequality we conclude that $\overline{\lambda} = |\beta|$ and the result follows. To prove the second part, observe that if Λ is a nonempty convex cone, namely, for any $\lambda \in \Lambda$ and $t \geq 0$ it holds that $t\lambda \in \Lambda$, we have that Ω is positive homogeneous. Indeed, making the change of variable $\lambda' = \lambda/|t|$ we see that $\Omega(t\beta|\Lambda) = |t|\Omega(\beta|\Lambda)$. Moreover, the above inequality, $\Omega(\beta|\Lambda) \geq \|\beta\|_1$, implies that if $\Omega(\beta|\Lambda) = 0$ then $\beta = 0$. The proof of the triangle inequality follows from the homogeneity and convexity of Ω, namely $\Omega(\alpha + \beta|\Lambda) = 2\Omega\left((\alpha + \beta)/2|\Lambda\right) \leq \Omega(\alpha|\Lambda) + \Omega(\beta|\Lambda)$.

Finally, note that $\Omega(\beta|\Lambda) \leq \omega\|\beta\|_1$ if and only if $\omega = \max\{\Omega(\beta|\Lambda) : \|\beta\|_1 = 1\}$. Since Ω is convex the maximum above is achieved at an extreme point of the ℓ_1 unit ball. ■

This proposition indicates a heuristic interpretation of the method (2.2): among all vectors β which have a fixed value of the ℓ_1 norm, the penalty function Ω will encourage those for which $|\beta| \in \Lambda$. Moreover, when $|\beta| \in \Lambda$ the function Ω reduces to the ℓ_1 norm and, so, the solution of problem (2.2) is expected to be sparse. The penalty function therefore will encourage certain desired sparsity patterns.

The last point can be better understood by looking at problem (2.1). For every solution $(\hat{\beta}, \hat{\lambda})$, the sparsity pattern of $\hat{\beta}$ is contained in the sparsity pattern of $\hat{\lambda}$, that is, the indices associated with nonzero components of $\hat{\beta}$ are a subset of those of $\hat{\lambda}$. Indeed, if $\hat{\lambda}_i = 0$ it must hold that $\hat{\beta}_i = 0$ as well, since the objective would diverge otherwise (because of the ratio β_i^2/λ_i). Therefore, if the set Λ favors certain sparse solutions of $\hat{\lambda}$, the same sparsity pattern will be reflected on $\hat{\beta}$. Moreover, the $\sum_{i \in \mathbb{N}_n} \lambda_i$ term appearing in the expression for $\Gamma(\beta, \lambda)$ favors sparse λ vectors. For example, a constraint of the form $\lambda_1 \geq \cdots \geq \lambda_n$ favors consecutive zeros at the end of λ and nonzeros everywhere else. This will lead to zeros at the terminal components of β as well. Thus, in many cases like this, it is easy to incorporate a convex constraint on λ, whereas it may not be possible to do the same with β.

Next, we note that a normalized version of the group Lasso penalty [31] is included in our setting as a special case. If, for some $k \in \mathbb{N}_n$, $\{J_\ell : \ell \in \mathbb{N}_k\}$ forms a partition of the index set $\mathbb{N}_n$, the corresponding group Lasso penalty is defined as

$$\Omega_{\mathrm{GL}}(\beta) = \sum_{\ell \in \mathbb{N}_k} \sqrt{|J_\ell|}\, \|\beta_{|J_\ell}\|_2, \tag{2.5}$$

where, for every $J \subseteq \mathbb{N}_n$, we use the notation $\beta_{|J} = (\beta_j : j \in J)$. It is an easy matter to verify that $\Omega_{\mathrm{GL}} = \Omega(\cdot|\Lambda)$ for $\Lambda = \{\lambda : \lambda \in \mathbb{R}^n_{++}, \lambda_j = \theta_\ell,\ j \in J_\ell,\ \ell \in \mathbb{N}_k,\ \theta_\ell > 0\}$.

The next proposition presents a useful construction which may be employed to generate new penalty functions from available ones. It is obtained by composing a set $\Theta \subseteq \mathbb{R}^k_{++}$ with a linear transformation, modeling the sum of the components of a vector, across the elements of a prescribed partition $\mathcal{P} = \{P_\ell : \ell \in \mathbb{N}_k\}$ of $\mathbb{N}_n$. To describe our result we introduce the *group average map* $A_{\mathcal{P}} : \mathbb{R}^n \to \mathbb{R}^k$ induced by $\mathcal{P}$. It is defined, for each $\beta \in \mathbb{R}^n$, as $A_{\mathcal{P}}(\beta) = (\|\beta_{|P_\ell}\|_1 : \ell \in \mathbb{N}_k)$.

Proposition 2.3. *If $\Theta \subseteq \mathbb{R}^k_{++}$, $\beta \in \mathbb{R}^n$ and $\mathcal{P}$ is a partition of $\mathbb{N}_n$ then*

$$\Omega(\beta|A_{\mathcal{P}}^{-1}(\Theta)) = \Omega(A_{\mathcal{P}}(\beta)|\Theta).$$

Proof. The idea of the proof depends on two basic observations. The first uses the set theoretic formula

$$A_{\mathcal{J}}^{-1}(\Theta) = \bigcup_{\theta \in \Theta} A_{\mathcal{J}}^{-1}(\theta).$$

From this decomposition we obtain that

$$\Omega(\beta | A_{\mathcal{J}}^{-1}(\Theta)) = \inf \left\{ \inf \left\{ \Gamma(\beta, \lambda) : \lambda \in A_{\mathcal{J}}^{-1}(\theta) \right\} : \theta \in \Theta \right\}. \tag{2.6}$$

Next, we write $\theta = (\theta_\ell : \ell \in \mathbb{N}_k) \in \Theta$ and decompose the inner infimum as the sum

$$\sum_{\ell \in \mathbb{N}_k} \inf \left\{ \frac{1}{2} \sum_{j \in J_\ell} \left(\frac{\beta_j^2}{\lambda_j} + \lambda_j \right) : \sum_{j \in J_\ell} \lambda_j = \theta_\ell, \lambda_j > 0, j \in J_\ell \right\}.$$

Now, the second essential step in the proof evaluates the infimum in the second sum by the Cauchy-Schwarz inequality to obtain that

$$\inf \left\{ \Gamma(\beta | \lambda) : \lambda \in A_{\mathcal{J}}^{-1}(\theta) \right\} = \sum_{\ell \in \mathbb{N}_k} \frac{1}{2} \left(\frac{\|\beta_{|J_\ell}\|_1^2}{\theta_\ell} + \theta_\ell \right).$$

We now substitute this formula into the right hand side of equation (2.6) to finish the proof. ■

When the set Λ is a nonempty convex cone, to emphasize that the function $\Omega(\cdot|\Lambda)$ is a norm we denoted it by $\|\cdot\|_\Lambda$. We end this section with the identification of the dual norm of $\|\cdot\|_\Lambda$, which is defined as

$$\|\beta\|_{*,\Lambda} = \max \left\{ \beta^\top u : u \in \mathbb{R}^n, \|u\|_\Lambda = 1 \right\}.$$

Proposition 2.4. *If Λ is a nonempty convex cone then there holds the equation*

$$\|\beta\|_{*,\Lambda} = \sup \left\{ \sqrt{\frac{\sum_{i \in \mathbb{N}_n} \lambda_i \beta_i^2}{\sum_{i \in \mathbb{N}_n} \lambda_i}} : \lambda \in \Lambda \right\}.$$

Proof. By definition, $\varphi = \|\beta\|_{*,\Lambda}$ is the smallest constant φ such that, for every $\lambda \in \Lambda$ and $u \in \mathbb{R}^n$, it holds that

$$\frac{\varphi}{2} \sum_{i \in \mathbb{N}_n} \left(\frac{u_i^2}{\lambda_i} + \lambda_i \right) - \beta^\top u \geq 0.$$

Minimizing the left hand side of this inequality for $u \in \mathbb{R}^n$ yields the equivalent inequality

$$\varphi^2 \geq \frac{\sum_{i \in \mathbb{N}_n} \lambda_i \beta_i^2}{\sum_{i \in \mathbb{N}_n} \lambda_i}.$$

Since this inequality holds for every $\lambda \in \Lambda$, the result follows by taking the supremum of the right hand side of the above inequality over this set. ■

The formula for the dual norm suggests that we introduce the set $\tilde{\Lambda} = \{\lambda : \lambda \in \Lambda, \sum_{i \in \mathbb{N}_n} \lambda_i = 1\}$. With this notation we see that the dual norm becomes

$$\|\beta\|_{*,\Lambda} = \sup \left\{ \sqrt{\sum_{i \in \mathbb{N}_n} \lambda_i \beta_i^2} : \lambda \in \tilde{\Lambda} \right\}.$$

Moreover, a direct computation yields an alternate form for the original norm given by the equation

$$\|\beta\|_{\Lambda} = \inf \left\{ \sqrt{\sum_{i \in \mathbb{N}_n} \frac{\beta_i^2}{\lambda_i}} : \lambda \in \tilde{\Lambda} \right\}.$$

3 Box penalty

We proceed to discuss some examples of the set $\Lambda \subseteq \mathbb{R}^n_{++}$ which may be used in the design of the penalty function $\Omega(\cdot|\Lambda)$.

The first example, which is presented in this section, corresponds to the prior knowledge that the magnitude of the components of the regression vector should be in some prescribed intervals. We choose $a = (a_i : i \in \mathbb{N}_n)$, $b = (b_i : i \in \mathbb{N}_n) \in \mathbb{R}^n$, $0 < a_i \leq b_i$ and define the corresponding box as $B[a, b] := \{(\lambda_i : i \in \mathbb{N}_n) : \lambda_i \in [a_i, b_i],\ i \in \mathbb{N}_n\}$. The theorem below establishes the form of the box penalty. To state our result, we define, for every $t \in \mathbb{R}$, the function $t_+ = \max(0, t)$.

Theorem 3.1. *We have that*

$$\Omega(\beta|B[a, b]) = \|\beta\|_1 + \sum_{i \in \mathbb{N}_n} \left(\frac{1}{2a_i}(a_i - |\beta_i|)_+^2 + \frac{1}{2b_i}(|\beta_i| - b_i)_+^2 \right).$$

Moreover, the components of the vector $\lambda(\beta) := \operatorname{argmin}\{\Gamma(\beta, \lambda) : \lambda \in B[a, b]\}$ *are given by the equations* $\lambda_i(\beta) = |\beta_i| + (a_i - |\beta_i|)_+ - (|\beta_i| - b)_+$, $i \in \mathbb{N}_n$.

Proof. Since $\Omega(\beta|B[a, b]) = \sum_{i \in \mathbb{N}_n} \Omega(\beta_i|[a_i, b_i])$ it suffices to establish the result in the case $n = 1$. We shall show that if $a, b, \beta \in \mathbb{R}$, $a \leq b$ then

$$\Omega(\beta|[a, b]) = |\beta| + \frac{1}{2a}(a - |\beta|)_+^2 + \frac{1}{2b}(|\beta| - b)_+^2. \tag{3.1}$$

Since both sides of the above equation are continuous functions of β it suffices to prove this equation for $\beta \in \mathbb{R}\backslash\{0\}$. In this case, the function $\Gamma(\beta, \cdot)$ is strictly convex, and so, has a unique minimum in $\mathbb{R}_{++}$ at $\lambda = |\beta|$, see also Figure 1-b. Moreover, if $|\beta| \leq a$ the minimum occurs at $\lambda = a$, whereas if $|\beta| \geq b$, it occurs at $\lambda = b$. This establishes the formula for $\lambda(\beta)$. Consequently, we have that

$$\Omega(\beta|[a, b]) = \begin{cases} |\beta|, & \text{if } |\beta| \in [a, b] \\ \frac{1}{2}\left(\frac{\beta^2}{a} + a\right), & \text{if } |\beta| < a \\ \frac{1}{2}\left(\frac{\beta^2}{b} + b\right), & \text{if } |\beta| > b. \end{cases}$$

Equation (3.1) now follows by a direct computation. ■

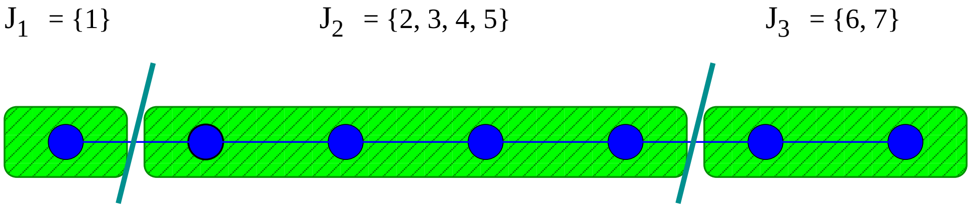


Figure 2: Partition of $\beta = (1.0732, -0.4872, 0.2961, -1.3692, 1.4731, -0.0073, -0.2133)$.

We also refer to [12, 24] for related penalty functions. Note that the function in equation (3.1) is a concatenation of two quadratic functions, connected together with a linear function. Thus, the box penalty will favor sparsity only for $a = 0$, case that is defined by a limiting argument.

4 Wedge penalty

In this section, we consider the case that the coordinates of the vector $\lambda \in \Lambda$ are ordered in a nonincreasing fashion. As we shall see, the corresponding penalty function favors regression vectors which are likewise nonincreasing.

We define the wedge

$$W = \{\lambda : \lambda = (\lambda_i : i \in \mathbb{N}_n) \in \mathbb{R}^n_{++}, \lambda_i \geq \lambda_{i+1},\ i \in \mathbb{N}_{n-1}\}.$$

Our next result describes the form of the penalty Ω in this case. To explain this result we require some preparation. We say that a partition $\mathcal{J} = \{J_\ell : \ell \in \mathbb{N}_k\}$ of $\mathbb{N}_n$ is *contiguous* if for all $i \in J_\ell, j \in J_{\ell+1}$, $\ell \in \mathbb{N}_{k-1}$, it holds that $i < j$. For example, if $n = 3$, partitions $\{\{1,2\},\{3\}\}$ and $\{\{1\},\{2\},\{3\}\}$ are contiguous but $\{\{1,3\},\{2\}\}$ is not.

Definition 4.1. *Given any two disjoint subsets $J, K \subseteq \mathbb{N}_n$ we define the region in $\mathbb{R}^n$*

$$Q_{J,K} = \left\{\beta : \beta \in \mathbb{R}^n, \frac{\|\beta_{|J}\|_2^2}{|J|} > \frac{\|\beta_{|K}\|_2^2}{|K|}\right\}. \tag{4.1}$$

Note that the boundary of this region is determined by the zero set of a homogeneous polynomial of degree two. We also need the following construction.

Definition 4.2. *For every $S \subseteq \mathbb{N}_{n-1}$ we set $k = |S|+1$ and label the elements of S in increasing order as $S = \{j_\ell : \ell \in \mathbb{N}_{k-1}\}$. We associate with the set S a contiguous partition of $\mathbb{N}_n$, given by $\mathcal{J}(S) = \{J_\ell : \ell \in \mathbb{N}_k\}$, where we define $J_\ell := [j_{\ell-1}+1, j_\ell] \cap \mathbb{N}_n$, $\ell \in \mathbb{N}_k$, and set $j_0 = 0$ and $j_k = n$.*

Figure 2 illustrates an example of a contiguous partition along with the set $\mathcal{J}(S)$.

A subset S of $\mathbb{N}_{n-1}$ also induces two regions in $\mathbb{R}^n$ which play a central role in the identification of the wedge penalty. First, we describe the region which "crosses over" the induced partition $\mathcal{J}(S)$. This is defined to be the set

$$O_S := \bigcap \left\{Q_{J_\ell, J_{\ell+1}} : \ell \in \mathbb{N}_{k-1}\right\}. \tag{4.2}$$

In other words, $\beta \in O_S$ if the average of the square of its components within each region J_ℓ strictly decreases with ℓ. The next region which is essential in our analysis is the "stays within" region, induced by the partition $\mathcal{J}(S)$. This region is defined as

$$I_S := \bigcap \left\{ \overline{Q}_{J_\ell, J_{\ell,q}} : q \in J_\ell, \ell \in \mathbb{N}_k \right\} \tag{4.3}$$

where $\overline{Q}$ denotes the closure of the set Q and we use the notation $J_{\ell,q} := \{j : j \in J_\ell, j \leq q\}$. In other words, all vectors β within this region have the property that, for every set $J_\ell \in \mathcal{J}(S)$, the average of the square of a first segment of components of β within this set is not greater than the average over J_ℓ. We note that if S is the empty set the above notation should be interpreted as $O_S = \mathbb{R}^n$ and

$$I_S = \bigcap \{ \overline{Q}_{\mathbb{N}_n, \mathbb{N}_q} : q \in \mathbb{N}_n \}.$$

From the cross-over and stay-within sets we define the region

$$P_S = O_S \cap I_S.$$

Alternatively, we shall describe below the set P_S in terms of two vectors induced by a vector $\beta \in \mathbb{R}^n$ and the set $S \subseteq \mathbb{N}_{n-1}$. These vectors play the role of the Lagrange multiplier and the minimizer λ for the wedge penalty in the theorem below.

Definition 4.3. *For every vector $\beta \in (\mathbb{R}\backslash\{0\})^n$ and every subset $S \subseteq \mathbb{N}_{n-1}$ we let $\mathcal{J}(S)$ be the induced contiguous partition of $\mathbb{N}_n$ and define two vectors $\zeta(\beta, S) \in \mathbb{R}_+^{n+1}$ and $\delta(\beta, S) \in \mathbb{R}_{++}^n$ by*

$$\zeta_q(\beta, S) = \begin{cases} 0, & \text{if } q \in S \cup \{0, n\}, \\ \\ |J_{\ell,q}| - |J_\ell| \frac{\|\beta_{|J_{\ell,q}}\|_2^2}{\|\beta_{|J_\ell}\|_2^2}, & \text{if } q \in J_\ell, \ell \in \mathbb{N}_k \end{cases}$$

and

$$\delta_q(\beta, S) = \frac{\|\beta_{|J_\ell}\|_2}{\sqrt{|J_\ell|}}, \; q \in J_\ell, \ell \in \mathbb{N}_k. \tag{4.4}$$

Note that the components of $\delta(\beta, S)$ are constant on each set J_ℓ, $\ell \in \mathbb{N}_k$.

Lemma 4.1. *For every $\beta \in (\mathbb{R}\backslash\{0\})^n$ and $S \subseteq \mathbb{N}_{k-1}$ we have that*

(a) $\beta \in P_S$ if and only if $\zeta(\beta, S) \geq 0$ *and* $\delta(\beta, S) \in \operatorname{int}(W)$*;*

(b) If $\delta(\beta, S_1) = \delta(\beta, S_2)$ and $\beta \in O_{S_1} \cap O_{S_2}$ then $S_1 = S_2$.

Proof. The first assertion follows directly from the definition of the requisite quantities. The proof of the second assertion is a direct consequence of the fact that the vector $\delta(\beta, S)$ is a constant on any element of the partition $\mathcal{J}(S)$ and strictly decreasing from one element to the next in that partition. ■

For the theorem below we introduce, for every $S \in \mathbb{N}_{n-1}$ the sets

$$U_S := P_S \cap (\mathbb{R}\backslash\{0\})^n.$$

We shall establishes not only that the collection of sets $\mathcal{U} := \{U_S : S \subseteq \mathbb{N}_{n-1}\}$ form a *partition* of $(\mathbb{R}\backslash\{0\})^n$, that is, their union is $(\mathbb{R}\backslash\{0\})^n$ and two distinct elements of $\mathcal{U}$ are disjoint, but also explicitly determine the wedge penalty on each element of $\mathcal{U}$.

Theorem 4.1. *The collection of sets $\mathcal{U} := \{U_S : S \subseteq \mathbb{N}_{n-1}\}$ form a partition of $(\mathbb{R}\backslash\{0\})^n$. For each $\beta \in (\mathbb{R}\backslash\{0\})^n$ there is a unique $S \subseteq \mathbb{N}_{n-1}$ such that $\beta \in \mathcal{U}_S$, and*

$$\|\beta\|_W = \sum_{\ell \in \mathbb{N}_k} \sqrt{|J_\ell|}\|\beta_{|J_\ell}\|_2, \tag{4.5}$$

where $k = |S| + 1$. Moreover, the components of the vector $\lambda(\beta) := \operatorname{argmin}\{\Gamma(\beta, \lambda) : \lambda \in W\}$ are given by the equations $\lambda_j(\beta) = \mu_\ell,\ j \in J_\ell,\ \ell \in \mathbb{N}_k$, where

$$\mu_\ell = \frac{\|\beta_{|J_\ell}\|_2}{\sqrt{|J_\ell|}}. \tag{4.6}$$

Proof. First, let us observe that there are $n-1$ inequality constraints defining W. It readily follows that all vectors in this constraint set are *regular*, in the sense of optimization theory, see [4, p. 279]. Hence, we can appeal to [4, Prop. 3.3.4, p. 316 and Prop. 3.3.6, p. 322], which state that $\lambda \in \mathbb{R}^n_{++}$ is a solution to the minimum problem determined by the wedge penalty, if and only if there exists a vector $\alpha = (\alpha_i : i \in \mathbb{N}_{n-1})$ with nonnegative components such that

$$-\frac{\beta_j^2}{\lambda_j^2} + 1 + \alpha_{j-1} - \alpha_j = 0, \quad j \in \mathbb{N}_n, \tag{4.7}$$

where we set $\alpha_0 = \alpha_n = 0$. Furthermore, the following complementary slackness conditions hold true

$$\alpha_j(\lambda_{j+1} - \lambda_j) = 0,\ j \in \mathbb{N}_{n-1}. \tag{4.8}$$

To unravel these equations, we let $\hat{S} := \{j : \lambda_j > \lambda_{j+1}, j \in \mathbb{N}_{n-1}\}$, which is the subset of indexes corresponding to the constraints that are not tight. When $k \geq 2$, we express this set in the form $\{j_\ell : \ell \in \mathbb{N}_{k-1}\}$ where $k = |\hat{S}| + 1$. As explained in Definition 4.2, the set $\hat{S}$ induces the partition $\mathcal{J}(\hat{S}) = \{J_\ell : \ell \in \mathbb{N}_k\}$ of $\mathbb{N}_n$. When $k = 1$ our notation should be interpreted to mean that $\hat{S}$ is empty and the partition $\mathcal{J}(\hat{S})$ consists only of $\mathbb{N}_n$. In this case, it is easy to solve equations (4.7) and (4.8). In fact, all components of the vector λ have a common value, say $\mu > 0$, and by summing both sides of equation (4.7) over $j \in \mathbb{N}_n$ we obtain that

$$\mu^2 = \frac{\|\beta\|_2^2}{n}.$$

Moreover, summing both sides of the same equation over $j \in \mathbb{N}_q$ we obtain that

$$\alpha_q = -\frac{\sum_{j \in \mathbb{N}_q} \beta_j^2}{\mu^2} + q$$

and, since $\alpha_q \geq 0$ we conclude that $\beta \in I_{\hat{S}} = P_{\hat{S}}$.

We now consider the case that $k \geq 2$. Hence, the vector λ has equal components on each subset J_ℓ, which we denote by μ_ℓ, $\ell \in \mathbb{N}_{k-1}$. The definition of the set $\hat{S}$ implies that the sequence $\{\mu_\ell : \ell \in \mathbb{N}_k\}$ is strictly decreasing and equation (4.8) implies that $\alpha_j = 0$, for every $j \in \hat{S}$. Summing both sides of equation (4.7) over $j \in J_\ell$ we obtain that

$$-\frac{1}{\mu_\ell^2} \sum_{j \in J_\ell} \beta_j^2 + |J_\ell| = 0 \tag{4.9}$$

from which equation (4.6) follows. Since the μ_ℓ are strictly decreasing, we conclude that $\beta \in O_{\hat{S}}$. Moreover, choosing $q \in J_\ell$ and summing both sides of equations (4.7) over $j \in J_{\ell,q}$ we obtain that

$$0 \leq \alpha_q = -\frac{\|\beta_{|J_{\ell,q}}\|_2^2}{\mu_\ell^2} + |J_{\ell,q}|$$

which implies that $\beta \in \overline{Q}_{J_\ell, J_{\ell,q}}$. Since this holds for every $q \in J_\ell$ and $\ell \in N_k$ we conclude that $\beta \in I_{\hat{S}}$ and therefore, it follows that $\beta \in U_S$.

In summary, we have shown that $\alpha = \zeta(\beta, \hat{S})$, $\lambda = \delta(\beta, \hat{S})$, and $\beta \in U_{\hat{S}}$. In particular, this implies that the collection of sets $\mathcal{U}$ covers $(\mathbb{R}\backslash\{0\})^n$. Next, we show that the elements of $\mathcal{U}$ are disjoint. To this end, we observe that, the computation described above can be *reversed*. That is to say, conversely for *any* $\hat{S} \subseteq \mathbb{N}_{n-1}$ and $\beta \in U_{\hat{S}}$ we conclude that $\delta(\beta, \hat{S})$ and $\zeta(\beta, \hat{S})$ solve the equations (4.7) and (4.8). Since the wedge penalty function is *strictly convex* we know that equations (4.7) and (4.8) have a unique solution. Now, if $\beta \in U_{S_1} \cap U_{S_2}$ then it must follow that $\delta(\beta, S_1) = \delta(\beta, S_2)$. Consequently, by part (b) in Lemma 4.1 we conclude that $S_1 = S_2$. ■

Note that the set S and the associated partition $\mathcal{J}$ appearing in the theorem is identified by examining the optimality conditions of the optimization problem (2.3) for $\Lambda = W$. There are 2^{n-1} possible partitions. Thus, for a given $\beta \in (\mathbb{R}\backslash\{0\})^n$, determining the corresponding partition is a challenging problem. We explain how to do this in Section 7.

An interesting property of the Wedge penalty, which is indicated by Theorem 4.1, is that it has the form of a group Lasso penalty as in equation (2.5), with groups not fixed *a-priori* but depending on the location of the vector β. The groups are the elements of the partition $\mathcal{J}$ and are identified by certain convex constraints on the vector β. For example, for $n = 2$ we obtain that $\Omega(\beta|W) = \|\beta\|_1$ if $|\beta_1| > |\beta_2|$ and $\Omega(\beta|W) = \sqrt{2}\|\beta\|_2$ otherwise. For $n = 3$, we have that

$$\Omega(\beta|W) = \begin{cases} \|\beta\|_1, & \text{if } |\beta_1| > |\beta_2| > |\beta_3| \qquad \mathcal{J} = \{\{1\},\{2\},\{3\}\} \\ \sqrt{2(\beta_1^2 + \beta_2^2)} + |\beta_3|, & \text{if } |\beta_1| \leq |\beta_2| \text{ and } \frac{\beta_1^2 + \beta_2^2}{2} > \beta_3^2 \qquad \mathcal{J} = \{\{1,2\},\{3\}\} \\ |\beta_1| + \sqrt{2(\beta_2^2 + \beta_3^2)}, & \text{if } |\beta_2| \leq |\beta_3| \text{ and } \beta_1^2 > \frac{\beta_2^2 + \beta_3^2}{2} \qquad \mathcal{J} = \{\{1\},\{2,3\}\} \\ \sqrt{3(\beta_1^2 + \beta_2^2 + \beta_3^2)}, & \text{otherwise} \qquad \mathcal{J} = \{\{1,2,3\}\} \end{cases}$$

where we have also displayed the partition $\mathcal{J}$ involved in each case. We also present a graphical representation of the corresponding unit ball in Figure 3-a. For comparison we also graphically

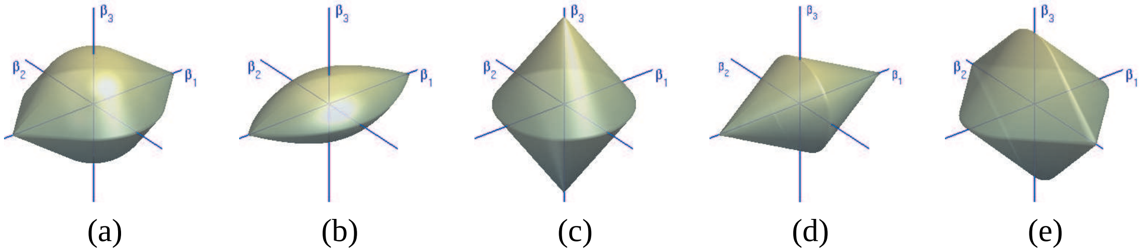


Figure 3: Unit ball of different penalty functions: (a) Wedge penalty $\Omega(\cdot|W)$; (b) hierarchical group Lasso; (c) group Lasso with groups $\{\{1,2\},\{3\}\}$; (d) group Lasso with groups $\{\{1\},\{2,3\}\}$; (e) the penalty $\Omega(\cdot|W^2)$.

display the unit ball for the hierarchical group Lasso with groups $\{1,2,3\},\{2,3\},\{3\}$ and two group Lasso in Figure 3-b,c,d, respectively.

The wedge may equivalently be expressed as the constraint that the difference vector $D^1(\lambda) := (\lambda_{j+1} - \lambda_j : j \in \mathbb{N}_{n-1})$ is less than or equal to zero. This alternative interpretation suggests the k-th order difference operator, which is given by the formula

$$D^k(\lambda) = \left(\lambda_{j+k} + \sum_{\ell \in \mathbb{N}_k} (-1)^\ell \binom{k}{\ell} \lambda_{j+k-\ell} : j \in \mathbb{N}_{n-k}\right)$$

and the corresponding k-th wedge

$$W^k := \{\lambda : \lambda \in \mathbb{R}^n_{++},\ D^k(\lambda) \geq 0\}. \tag{4.10}$$

The associated penalty $\Omega(\cdot|W^k)$ encourages vectors whose sparsity pattern is concentrated on at most k different contiguous regions. Note that W^1 is not the wedge W considered earlier. Moreover, the 2-wedge includes vectors which have a convex "profile" and whose sparsity pattern is concentrated either on the first elements of the vector, on the last, or on both.

5 Graph penalty

In this section we present an extension of the wedge set which is inspired by previous work on the group Lasso estimator with hierarchically overlapping groups [32]. It models vectors whose magnitude is ordered according to a graphical structure.

Let $G = (V,E)$ be a directed graph, where V is the set of n vertices in the graph and $E \subseteq V \times V$ is the edge set, whose cardinality is denoted by m. If $(v,w) \in E$ we say that there is a directed edge from vertex v to vertex w. The graph is identified by the $m \times n$ *incidence matrix*, which we define as

$$A_{e,v} = \begin{cases} 1, & \text{if } e = (v,w) \in E,\ w \in V, \\ -1, & \text{if } e = (w,v) \in E,\ w \in V, \\ 0, & \text{otherwise.} \end{cases}$$

We consider the penalty $\|\cdot\|_{\Lambda_G}$ for the convex cone $\Lambda_G = \{\lambda : \lambda \in \mathbb{R}^n_{++}, A\lambda \geq 0\}$ and assume, from now on, that G is acyclic (DAG), that is, G has no directed loops. In particular, this implies that, if $(v, w) \in E$ then $(w, v) \notin E$. The wedge penalty described above is a special case of the graph penalty corresponding to a line graph. Let us now discuss some aspects of the graph penalty for an arbitrary DAG. As we shall see, our remarks lead to an explicit form of the graph penalty when G is a tree.

If $(v, w) \in E$ we say that vertex w is a child of vertex v and v is a parent of w. For every vertex $v \in V$, we let $C(v)$ and $P(v)$ be the set of children and parents of v, respectively. When G is a tree, $P(v)$ is the empty set if v is the root node and otherwise $P(v)$ consists of only one element, the parent of v, which we denote by $p(v)$.

Let $D(v)$ be the set of descendants of v, that is, the set of vertices which are connected to v by a directed path starting in v, and let $A(v)$ be the set of ancestors of v, that is, the set of vertices from which a directed path leads to v. We use the convention that $v \in D(v)$ and $v \notin A(v)$.

Every connected subset $V' \subseteq V$ induces a subgraph of G which is also a DAG. If V_1 and V_2 are disjoint connected subsets of V, we say that they are connected if there is at least one edge connecting a pair of vertices in V_1 and V_2, in either one or the other direction. Moreover, we say that V_2 is below V_1 — written $V_2 \Downarrow V_1$ — if V_1 and V_2 are connected and every edge connecting them departs from a node of V_1.

Definition 5.1. *Let G be a DAG. We say that $C \subseteq E$ is a cut of G if it induces a partition $\mathcal{V}(C) = \{V_\ell : \ell \in \mathbb{N}_k\}$ of the vertex set V such that $(v, w) \in C$ if and only if vertices v and w belong to two different elements of the partition.*

In other words, a cut separates a connected graph in two or more connected components such that every pair of vertices corresponding to a disconnected edge, that is an element of C, are in two different components. We also denote by $\mathcal{C}(G)$ the set of cuts of G, and by $D_\ell(v)$ the set of descendants of v within set V_ℓ, for every $v \in V_\ell$ and $\ell \in \mathbb{N}_k$.

Next, for every $C \in \mathcal{C}(G)$, we define the regions in $\mathbb{R}^n$ by the equations

$$O_C = \bigcap \{Q_{V_1,V_2} : \ V_1, V_2 \in \mathcal{V}(C), V_2 \Downarrow V_1\} \tag{5.1}$$

and

$$I_C = \bigcap \left\{\overline{Q}_{D_\ell(v),V_\ell} : \ell \in \mathbb{N}_k, v \in V_\ell\right\}. \tag{5.2}$$

These sets are the graph equivalent of the sets defined by equations (4.2) and (4.3) in the special case of the wedge penalty in Section 4. We also set $P_C = O_C \cap I_C$.

Moreover, for every $C \in \mathcal{C}(G)$, we define the sets

$$U_C := P_C \bigcap (\mathbb{R}\backslash\{0\})^n.$$

As of yet, we cannot extend Theorem 4.1 to the case of an arbitrary DAG. However, we can accomplish this when G is a tree.

Lemma 5.1. *Let $G = (V, E)$ be a tree, let A be the associated incidence matrix and let $z = (z_v : v \in V) \in \mathbb{R}^n$. The following facts are equivalent:*

(a) *For every $v \in V$ it holds that*

$$\sum_{u \in D(v)} z_u \geq 0.$$

(b) *The linear system $A^\top \alpha = -z$ admits a non-negative solution for $\alpha = (\alpha_e : e \in E) \in \mathbb{R}^m$.*

Proof. The incident matrix of a tree has the property that, for every $v \in V$ and $e \in E$,

$$\sum_{u \in D(v)} A_{eu} = -\delta_{e,(p(v),v)} \tag{5.3}$$

where, for every $e, e' \in E$, $\delta_{e,e'} = 1$ if $e = e'$ and zero otherwise. The linear system in (b) can be written componentwise as

$$\sum_{e \in E} A_{eu} \alpha_e = -z_u.$$

Summing both sides of this equation over $u \in D(v)$ and using equation (5.3), we obtain the equivalent equations

$$\alpha_{(p(v),v)} = \sum_{u \in D(v)} z_u.$$

The result follows. ■

Definition 5.2. *Let $G = (V, E)$ be a DAG. For every vector $\beta \in (\mathbb{R} \backslash \{0\})^n$ and every cut $C \in \mathcal{C}(G)$ we let $\mathcal{V}(C) = \{V_\ell : \ell \in \mathbb{N}_k\}$, $k \in \mathbb{N}_n$ be the partition of V induced by C, and define two vectors $\zeta(\beta, C) \in \mathbb{R}_+^{n-1}$ and $\delta(\beta, C) \in \mathbb{R}_{++}^n$. The components of $\zeta(\beta, C)$ are given as*

$$\zeta_e(\beta, C) = \begin{cases} 0, & \text{if } e \in C, \\ \\ |V_\ell| \frac{\|\beta_{|D_\ell(u)}\|_2^2}{\|\beta_{|V_\ell}\|_2^2} - |D_\ell(u)|, & \text{if } e = (u, v), u \in V_\ell, v \in D_\ell(u),\ \ell \in \mathbb{N}_k \end{cases}$$

whereas the components of $\delta(\beta, C)$ are given by

$$\delta_v(\beta, C) = \frac{\|\beta_{|V_\ell}\|_2}{\sqrt{|V_\ell|}},\ v \in V_\ell,\ \ell \in \mathbb{N}_k. \tag{5.4}$$

Note that the notation we adopt in this definition differs from that used in the case of line graph, given in Definition 4.3. However, Definition 5.2 leads to a more appropriate presentation of our results for a tree.

Proposition 5.1. *Let $G = (V, E)$ be a tree and A the associated incidence matrix. For every $\beta \in (\mathbb{R} \backslash \{0\})^n$ and every cut $C \in \mathcal{C}(G)$ we have that*

(a) $\beta \in P_C$ if and only if $\zeta(\beta, C) \geq 0$, *$A\delta(\beta, C) \geq 0$ and $\delta_v(\beta, C) > \delta_w(\beta, C)$, for all $v \in V_1, w \in V_2$, $(v, w) \in E$, $V_1, V_2 \in \mathcal{V}(C)$;*

(b) *If $\delta(\beta, C_1) = \delta(\beta, C_2)$ and $\beta \in O_{C_1} \cap O_{C_2}$ then $C_1 = C_2$.*

Proof. We immediately see that $\beta \in O_C$ if and only if $A\delta(\beta, C) \geq 0$ and $\delta_v(\beta, C) > \delta_w(\beta, C)$ for all $v \in V_1, w \in V_2$, $(v, w) \in E$, $V_1, V_2 \in \mathcal{V}(C)$. Moreover, by applying Lemma 5.1 on each element V_ℓ of the partition induced by C and choosing $z = (|V_\ell| \frac{\beta_v^2}{\|\beta_{|V_\ell}\|_2^2} - 1 : v \in V_\ell)$, we conclude that $\zeta(\beta, C) \geq 0$ if and only if $\beta \in I_C$. This proves the first assertion.

The proof of the second assertion is a direct consequence of the fact that the vector $\delta(\beta, C)$ is a constant on any element of the partition $\mathcal{V}(C)$ and strictly decreasing from one element to the next in that partition. ■

Theorem 5.1. *Let $G = (V, E)$ be a tree. The collection of sets $\mathcal{U} := \{U_C : C \in \mathcal{C}(G)\}$ form a partition of $(\mathbb{R}\backslash\{0\})^n$. Moreover, for every $\beta \in (\mathbb{R}\backslash\{0\})^n$ there is a unique $C \in \mathcal{C}(G)$ such that*

$$\|\beta\|_{\Lambda_G} = \sum_{V_\ell \in \mathcal{V}(\mathcal{C})} \sqrt{|V_\ell|}\|\beta_{|V_\ell}\|_2 \tag{5.5}$$

and the vector $\lambda(\beta) = (\lambda_v(\beta) : v \in V)$ has components given by $\lambda_v(\beta) = \mu_\ell,\ v \in V_\ell$, $\ell \in \mathbb{N}_k$, where

$$\mu_\ell = \sqrt{\frac{1}{n_\ell}\sum_{w \in V_\ell} \beta_w^2}. \tag{5.6}$$

Proof. The proof of this theorem proceeds in a fashion similar to that of Theorem 4.1. In this regard, Lemma 5.1 is crucial. By KKT theory (see e.g. [4, Theorems 3.3.4,3.3.7]), λ is an optimal solution of the graph penalty if and only if there exists $\alpha \geq 0$ such that, for every $v \in V$

$$-\frac{\beta_v^2}{\lambda_v^2} + 1 - \sum_{e \in E} \alpha_e A_{ev} = 0$$

and the following complementary conditions hold true

$$\alpha_{(v,w)}(\lambda_w - \lambda_v) = 0,\ v \in V, w \in C(v). \tag{5.7}$$

We rewrite the first equation as

$$\alpha_{(p(v),v)} - \sum_{w \in C(v)} \alpha_{(v,w)} = \frac{\beta_v^2}{\lambda_v^2} - 1. \tag{5.8}$$

Now, if $\lambda \in \Lambda_G$ solves equations (5.7) and (5.8), then it induces a cut $C \subset E$ and a corresponding partition $\mathcal{V}(C) = \{V_\ell : \ell \in \mathbb{N}_k\}$ of V such that $\lambda_v = \mu_\ell$ for every $v \in V_\ell$. That is, $\lambda_v = \lambda_w$ for every $v, w \in V_\ell$, $\ell \in \mathbb{N}_k$, and $\alpha_e = 0$ for every $e \in C$. Therefore, summing equations (5.8) for $v \in V_\ell$ we get that

$$\mu_\ell = \frac{\|\beta_{|V_\ell}\|_2}{\sqrt{|V_\ell|}}.$$

Moreover, since $\mu_\ell > \mu_q$, if $V_q \Downarrow V_\ell$ we see that $\beta \in O_C$. Next, for every $\ell \in \mathbb{N}_k$ and $u \in V_\ell$ we sum both sides of equation (5.8) for $v \in D_\ell(u)$ to obtain that

$$\alpha_{(p(u),u)} = \frac{\|\beta_{|D_\ell(u)}\|_2^2}{\mu_\ell^2} - |D_\ell(u)|. \tag{5.9}$$

We see that $\beta \in I_C$ and conclude that $\beta \in U_C$.

In summary we have shown that the collection of sets $\mathcal{U}$ cover $(\mathbb{R}\backslash\{0\})^n$. Next, we show that the elements of $\mathcal{U}$ are disjoint. To this end, we observe that, the computation described above can be *reversed*. That is to say, conversely for *any* partition $C = \{V_i : i \in \mathbb{N}_k\}$ of V and $\beta \in U_C$ we conclude by Proposition 5.1 that the vectors $\delta(\beta, C)$ and $\zeta(\beta, C)$ solves the KKT optimality conditions. Since this solution is unique if $\beta \in U_{C_1} \cap U_{C_2}$ then it must follow that $\delta(\beta, C_1) = \delta(\beta, C_2)$, which implies that $C_1 = C_2$. ■

Theorems 4.1 and 5.1 fall into the category of a set $\Lambda \subseteq \mathbb{R}^n$ chosen in the form of a polyhedral cone, that is

$$\Lambda = \{\lambda : \lambda \in \mathbb{R}^n, A\lambda \geq 0\}$$

where A is an $m \times n$ matrix. Furthermore, in the line graph of Theorem 4.1 and also the extension in Theorem 5.1 the matrix A only has elements which are $-1, 1$ or 0. These two examples that we considered led to explicit description of the norm $\|\cdot\|_\Lambda$. However, there are seemingly simple cases of a matrix A of this type where the explicit computation of the norm $\|\cdot\|_\Lambda$ seem formidable, if not impossible. For example, if $m = 2$, $n = 4$ and

$$A = \quad \begin{bmatrix} -1 & -1 & 1 & 0 \\ 0 & -1 & -1 & 1 \end{bmatrix}$$

we are led by KKT to a system of equations that, in the case of two active constraints, that is, $A\lambda = 0$, are the common zeros of two *fourth order* polynomials in the vector $\lambda \in \mathbb{R}^2$.

6 Duality

In this section, we comment on the utility of the class of penalty functions considered in this paper, which is fundamentally based on their construction as constrained infimum of quadratic functions. To emphasize this point both theoretically and computationally, we discuss the conversion of the regularization variational problem over $\beta \in \mathbb{R}^n$, namely

$$\mathcal{E}(\Lambda) = \inf\{E(\beta, \lambda) : \beta \in \mathbb{R}^n, \lambda \in \Lambda\} \tag{6.1}$$

where

$$E(\beta, \lambda) := \|y - X\beta\|_2^2 + 2\rho\Gamma(\beta, \lambda),$$

into a variational problem over $\lambda \in \Lambda$.

To explain what we have in mind, we introduce the following definition.

Definition 6.1. *For every $\lambda \in \mathbb{R}^n_+$, we define the vector $\beta(\lambda) \in \mathbb{R}^n$ as*

$$\beta(\lambda) = \operatorname{diag}(\lambda) M(\lambda) X^\top y$$

where $M(\lambda) := (\operatorname{diag}(\lambda) X^\top X + \rho I)^{-1}$.

Note that $\beta(\lambda) = \operatorname{argmin}\{E(\beta, \lambda) : \beta \in \mathbb{R}^n\}$.

Theorem 6.1. *For $\rho > 0$, $y \in \mathbb{R}^m$, any $m \times n$ matrix X and any nonempty convex set Λ we have that*

$$\mathcal{E}(\Lambda) = \min\left\{\rho y^\top \left(X\mathrm{diag}(\lambda)X^\top + \rho I\right)^{-1} y + \rho \mathrm{tr}(\mathrm{diag}(\lambda)) : \lambda \in \overline{\Lambda} \cap \mathbb{R}^n_+\right\} \tag{6.2}$$

Moreover, if $\hat{\lambda}$ is a solution to this problem, then $\beta(\hat{\lambda})$ is a solution to problem (6.1).

Proof. We substitute the formula for $\Omega(\beta|\Lambda)$ into the right hand side of equation (6.1) to obtain that

$$\mathcal{E}(\Lambda) = \inf\left\{H(\lambda) : \lambda \in \Lambda\right\} \tag{6.3}$$

where we define

$$H(\lambda) = \min\left\{E(\beta, \lambda) : \beta \in \mathbb{R}^n\right\}.$$

A straightforward computation yields that

$$H(\lambda) = \rho y^\top \left(X\mathrm{diag}(\lambda)X^\top + \rho I\right)^{-1} y + \rho \mathrm{tr}(\mathrm{diag}(\lambda)).$$

Since $H(\lambda) \geq \rho \mathrm{tr}(\mathrm{diag}(\lambda))$, we conclude that any minimizing sequence for the optimization problem on the right hand side of equation (6.3) must have a subsequence which converges. These remarks confirm equation (6.2).

We now prove the second claim. For $\lambda \in \mathbb{R}^n_{++}$ a direct computation confirms that

$$\Gamma(\beta(\lambda), \lambda) = \frac{1}{2}\left(y^\top X M(\lambda)\mathrm{diag}(\lambda)M(\lambda)X^\top y + \mathrm{tr}(\mathrm{diag}(\lambda))\right).$$

Note that the right hand side of this equation provides a continuous extension of the left hand side to $\lambda \in \mathbb{R}^n_+$. For notational simplicity, we still use the left hand side to denote this *continuous extension*.

By a limiting argument, we conclude, for every $\lambda \in \overline{\Lambda}$, that

$$\Omega(\beta(\lambda)|\Lambda) \leq \Gamma(\beta(\lambda), \lambda). \tag{6.4}$$

We are now ready to complete the proof of the theorem. Let $\hat{\lambda}$ be a solution for the optimization problem (6.2). By definition, it holds, for any $\beta \in \mathbb{R}^n$ and $\lambda \in \overline{\Lambda}$, that

$$\|y - X\beta(\hat{\lambda})\|_2^2 + 2\rho\Gamma(\beta(\hat{\lambda}), \hat{\lambda}) = H(\hat{\lambda}) \leq H(\lambda) \leq \|y - X\beta\|_2^2 + 2\rho\Gamma(\beta, \lambda).$$

Combining this inequality with inequality (6.4) evaluated at $\lambda = \hat{\lambda}$, we conclude that

$$\|y - X\beta(\hat{\lambda})\|_2^2 + 2\rho\Omega(\beta(\hat{\lambda})|\Lambda) \leq \|y - X\beta\|_2^2 + 2\rho\Gamma(\beta, \lambda)$$

from which the result follows. ■

An important consequence of the above theorem is a method to find a solution $\hat{\beta}$ to the optimization problem (6.1) from a solution to the optimization problem (6.2). We illustrate this idea in the case that $X = I$.

Corollary 6.1. *It holds that*

$$\min\left\{\|\beta - y\|_2^2 + 2\rho\Omega(\beta|\Lambda) : \beta \in \mathbb{R}^n\right\} = \rho \min\left\{\sum_{i\in\mathbb{N}_n} \frac{y_i^2}{\lambda_i + \rho} + \lambda_i : \lambda \in \overline{\Lambda}\right\}. \tag{6.5}$$

Moreover, if $\hat{\lambda}$ is a solution of the right optimization problem then the vector $\beta(\hat{\lambda}) = (\beta_i(\hat{\lambda}) : i \in \mathbb{N}_n)$, whose components are defined for $i \in \mathbb{N}_n$ as

$$\beta_i(\hat{\lambda}) = \frac{\hat{\lambda}_i y_i}{\hat{\lambda}_i + \rho} \tag{6.6}$$

is a solution of the left optimization problem problem.

We further discuss two choices of the set Λ in which we are able to solve problem (6.5) analytically. The first case we consider is $\Lambda = \mathbb{R}_{++}^n$, which corresponds to the Lasso penalty. It is an easy matter to see that $\hat{\lambda} = (|y| - \rho)_+$ and the corresponding regression vector is obtained by the well-known "soft thresolding" formula $\beta(\hat{\lambda}) = (|y| - \rho)_+ \mathrm{sign}(y)$. The second case is the Wedge penalty. We find that the solution of the optimization problem in the right hand side of equation (6.5) is $\hat{\lambda} = (\lambda(y) - \rho)_+$, where $\lambda(y)$ is given in Theorem 4.1. Finally, we note that Corollary 6.1 and the example following it extend to the case that $X^\top X = I$ by replacing throughout the vector y by the vector $X^\top y$. In the statistical literature this setting is referred to as orthogonal design.

7 Optimization method

In this section, we address the issue of implementing the learning method (2.2) numerically.

Since the penalty function $\Omega(\cdot|\Lambda)$ is constructed as the infimum of a family of quadratic regularizers, the optimization problem (2.2) reduces to a simultaneous minimization over the vectors β and λ. For a fixed $\lambda \in \Lambda$, the minimum over $\beta \in \mathbb{R}^n$ is a standard Tikhonov regularization and can be solved directly in terms of a matrix inversion. For a fixed β, the minimization over $\lambda \in \Lambda$ requires computing the penalty function (2.3). These observations naturally suggests an alternating minimization algorithm, which has already been considered in special cases in [1]. To describe our algorithm we choose $\epsilon > 0$ and introduce the mapping $\phi^\epsilon : \mathbb{R}^n \to \mathbb{R}_{++}^n$, whose i-th coordinate at $\beta \in \mathbb{R}^n$ is given by

$$\phi_i^\epsilon(\beta) = \sqrt{\beta_i^2 + \epsilon}.$$

For $\beta \in (\mathbb{R}\backslash\{0\})^n$, we also let $\lambda(\beta) = \mathrm{argmin}\{\Gamma(\beta, \lambda) : \lambda \in \Lambda\}$.

The alternating minimization algorithm is defined as follows: choose $\lambda_0 \in \Lambda$ and, for $k \in \mathbb{N}$, define the iterates

$$\begin{aligned} \beta^k &= \beta(\lambda^{k-1}) & (7.1)\\ \lambda^k &= \lambda(\phi^\epsilon(\beta^k)). & (7.2) \end{aligned}$$

The following theorem establishes convergence of this algorithm.

Theorem 7.1. *If the set Λ is admissible in the sense of Definition 2.1, then the iterations (7.1)–(7.2) converges to a vector $\gamma(\epsilon)$ such that*

$$\gamma(\epsilon) = \operatorname{argmin}\left\{\|y - X\beta\|_2^2 + 2\rho\Omega(\phi^\epsilon(\beta)|\Lambda) : \beta \in \mathbb{R}^n\right\}.$$

Moreover, any convergent subsequence of the sequence $\{\gamma\left(\frac{1}{\ell}\right) : \ell \in \mathbb{N}\}$ converges to a solution of the optimization problem (2.2).

Proof. We divide the proof into several steps. To this end, we define

$$E_\epsilon(\beta, \lambda) := \|y - X\beta\|_2^2 + 2\rho\Gamma(\phi^\epsilon(\beta), \lambda)$$

and note that $\beta(\lambda) = \operatorname{argmin}\{E_\epsilon(\alpha, \lambda) : \alpha \in \mathbb{R}^n\}$.

Step 1. We define two sequences, $\theta_k = E_\epsilon(\beta^k, \lambda^{k-1})$ and $\nu_k = E_\epsilon(\beta^k, \lambda^k)$ and observe, for any $k \geq 2$, that

$$\nu_k \leq \theta_k \leq \nu_{k-1}. \tag{7.3}$$

These inequalities follow directly from the definition of the alternating algorithm, see equations (7.1) and (7.2).

Step 2. We define the compact set $B = \{\beta : \beta \in \mathbb{R}^n, \|\beta\|_1 \leq \theta_1\}$. From the first inequality in Proposition 2.2 and inequality (7.3) we conclude, for every $k \in \mathbb{N}$, that $\beta^k \in B$.

Step 3. We define the function $g : \mathbb{R}^n \to \mathbb{R}$ at $\beta \in \mathbb{R}^n$ as

$$g(\beta) = \min\left\{E_\epsilon(\alpha, \lambda(\phi^\epsilon(\beta))) : \alpha \in \mathbb{R}^n\right\}.$$

We claim that g is continuous on B. In fact, there exists a constant $\kappa > 0$ such that, for every $\gamma^1, \gamma^2 \in B$, it holds that

$$|g(\gamma^1) - g(\gamma^2)| \leq \kappa\|\lambda(\phi^\epsilon(\gamma^1)) - \lambda(\phi^\epsilon(\gamma^2))\|_\infty. \tag{7.4}$$

The essential ingredient in the proof of this inequality is the fact that there exists constant $\overline{a}$ and $\overline{b}$ such that, for all $\beta \in B$, $\lambda(\phi^\epsilon(\beta)) \in [\overline{a}, \overline{b}]^n$. This follows from the inequalities developed in the proof of Proposition 2.1.

Step 4. By step 2, there exists a subsequence $\{\beta^{k_\ell} : \ell \in \mathbb{N}\}$ which converges to $\tilde{\beta} \in B$ and, for all $\beta \in \mathbb{R}^n$ and $\lambda \in \Lambda$, it holds that

$$E_\epsilon(\tilde{\beta}, \lambda(\phi^\epsilon(\tilde{\beta}))) \leq E_\epsilon(\beta, \lambda(\phi^\epsilon(\tilde{\beta}))), \quad E_\epsilon(\tilde{\beta}, \lambda(\phi^\epsilon(\tilde{\beta}))) \leq E_\epsilon(\tilde{\beta}, \lambda). \tag{7.5}$$

Indeed, from step 1 we conclude that there exists $\psi \in \mathbb{R}_{++}$ such that

$$\lim_{k\to\infty} \theta_k = \lim_{k\to\infty} \nu_k = \psi.$$

Since, by Proposition 2.1 $\lambda(\beta)$ is continuous for $\beta \in (\mathbb{R}\backslash\{0\})^n$, we obtain that

$$\lim_{\ell\to\infty} \lambda^{k_\ell} = \lambda(\phi^\epsilon(\tilde{\beta})).$$

By the definition of the alternating algorithm, we have, for all $\beta \in \mathbb{R}^n$ and $\lambda \in \Lambda$, that

$$\theta_{k+1} = E_\epsilon(\beta^{k+1}, \lambda^k) \leq E_\epsilon(\beta, \lambda^k), \qquad \nu_k = E_\epsilon(\beta^k, \lambda^k) \leq E_\epsilon(\beta^k, \lambda).$$

```
**Initialization:** $k \leftarrow 0$
**Input:** $\beta \in \mathbb{R}^n$;   **Output:** $J_1, \dots, J_k$
**for** $t = 1$ **to** $n$ **do**
    $J_{k+1} \leftarrow \{t\}$;
    $k \leftarrow k + 1$
    **while** $k > 1$ **and** $\frac{\|\beta_{|J_{k-1}}\|_2}{\sqrt{|J_{k-1}|}} \leq \frac{\|\beta_{|J_k}\|_2}{\sqrt{|J_k|}}$
        $J_{k-1} \leftarrow J_{k-1} \cup J_k$
        $k \leftarrow k - 1$
    **end**
**end**
```

Figure 4: Iterative algorithm to compute the wedge penalty

From this inequality we obtain, passing to limit, inequalities (7.5).

Step 5. The vector $(\tilde{\beta}, \lambda(\phi^\epsilon(\tilde{\beta})))$ is a stationary point. Indeed, since Λ is admissible, by step 3, $\lambda(\phi^\epsilon(\tilde{\beta}) \in \mathrm{int}(\Lambda)$. Therefore, since E_ϵ is continuously differentiable this claim follows from step 4.

Step 6. The alternating algorithm converges. This claim follows from the fact that E_ϵ is strictly convex. Hence, E_ϵ has a unique global minimum in $\mathbb{R}^n \times \Lambda$, which in virtue of inequalities (7.5) is attained at $(\tilde{\beta}, \lambda(\phi^\epsilon(\tilde{\beta})))$.

The last claim in the theorem follows from the fact that the set $\{\gamma(\epsilon) : \epsilon > 0\}$ is bounded and the function $\lambda(\beta)$ is continuous. ∎

The most challenging step in the alternating algorithm is the computation of the vector λ^k. Fortunately, if Λ is a second order cone, problem (2.3) defining the penalty function $\Omega(\cdot|\Lambda)$ may be reformulated as a second order cone program (SOCP), see e.g. [6]. To see this, we introduce an additional variable $t \in \mathbb{R}^n$ and note that

$$\Omega(\beta|\Lambda) = \min \left\{ \sum_{i \in \mathbb{N}_n} t_i + \lambda_i : \|(2\beta_i, t_i - \lambda_i)\|_2 \leq t_i + \lambda_i, t_i \geq 0,\ i \in \mathbb{N}_n,\ \lambda \in \Lambda \right\}.$$

In particular, the examples discussed in Sections 4 and 5, the set Λ is formed by linear constraints and, so, problem (2.3) is an SOCP. We may then use available tool-boxes to compute the solution of this problem. However, in special cases the computation of the penalty function may be significantly facilitated by using available analytical formulas. Here, for simplicity we describe how to do this in the case of the wedge penalty. For this purpose we say that a vector $\beta \in \mathbb{R}^n$ is admissible if, for every $k \in \mathbb{N}_n$, it holds that $\|\beta_{|\mathbb{N}_k}\|_2/\sqrt{k} \leq \|\beta\|_2/\sqrt{n}$.

The proof of the next lemma is straightforward and we do not elaborate on the details.

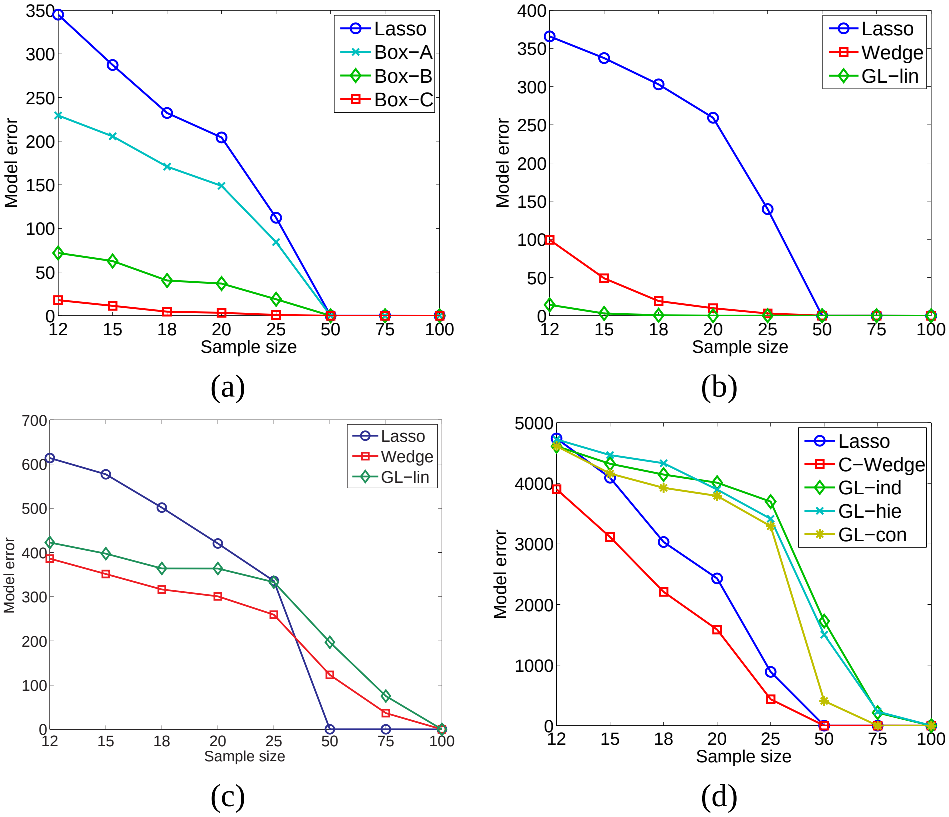
Lemma 7.1. *If $\beta \in \mathbb{R}^n$ and $\delta \in \mathbb{R}^p$ are admissible and $\|\beta\|_2/\sqrt{n} \leq \|\delta\|_2/\sqrt{p}$ then (β, δ) is admissible.*

The iterative algorithm presented in Figure 4 can be used to find the partition $\mathcal{J} = \{J_\ell : \ell \in \mathbb{N}_k\}$ and, so, the vector $\lambda(\beta)$ described in Theorem 4.1. The algorithm processes the components of vector β in a sequential manner. Initially, the first component forms the only set in the partition. After the generic iteration $t-1$, where the partition is composed of k sets, the index of the next components, t, is put in a new set J_{k+1}. Two cases can occur: the means of the squares of the sets are in strict descending order, or this order is violated by the last set. The latter is the only case that requires further action, so the algorithm merges the last two sets and repeats until the sets in the partition are fully ordered. Note that, since the only operation performed by the algorithm is the merge of admissible sets, Lemma 7.1 ensures that after each step t the current partition satisfies the "stay within" conditions $\frac{\|\beta_{|K}\|_2}{\sqrt{k}} > \frac{\|\beta_{|J_\ell \setminus K}\|_2}{\sqrt{|J_\ell|-k}}$, for every $\ell \in \mathbb{N}_k$ and every subset $K \subset J_\ell$ formed by the first $k < |J_\ell|$ elements of J_ℓ. Moreover, the *while* loop ensures that after each step the current partition satisfies, for every $\ell \in \mathbb{N}_{k-1}$, the "cross over" conditions $\|\beta_{|J_\ell}\|_2\sqrt{|J_\ell|} > \|\beta_{|J_{\ell+1}}\|_2\sqrt{|J_{\ell+1}|}$. Thus, the output of the algorithm is the partition $\mathcal{J}$ defined in Theorem 4.1. In the actual implementation of the algorithm, the means of squares of each set can be saved. This allows us to compute the mean of squares of a merged set as a weighted mean, which is a constant time operation. Since there are $n-1$ consecutive terms in total, this is also the maximum number of merges that the algorithm can perform. Each merge requires exactly one additional test, so we can conclude that the running time of the algorithm is linear.

8 Numerical simulations

In this section we present some numerical simulations with the proposed method. For simplicity, we consider data generated noiselessly from $y = X\beta^*$, where $\beta^* \in \mathbb{R}^{100}$ is the true underlying regression vector, and X is an $m \times 100$ input matrix, m being the sample size. The elements of X are generated i.i.d. from the standard normal distribution, and the columns of X are then normalized such that their ℓ_2 norm is 1. Since we consider the noiseless case, we solve the interpolation problem $\min\{\Omega(\beta) : y = X\beta\}$, for different choices of the penalty function Ω. In practice, (2.2) is solved for a tiny value of the parameter, for example, $\rho = 10^{-8}$, which we found to be sufficient to ensure that the error term in (2.2) is negligible at the minimum. All experiments were repeated 50 times, generating each time a new matrix X. In the figures we report the average of the model error of the vector $\hat{\beta}$ learned by each method, as a function of the sample size m. The former is defined as $\mathrm{ME}(\hat{\beta}) = \mathbb{E}[\|\hat{\beta} - \beta^*\|_2^2]/\mathbb{E}[\|\beta^*\|_2^2]$. In the following, we discuss a series of experiments, corresponding to different choices for the model vector β^* and its sparsity pattern. In all experiments, we solved the optimization problem (2.2) with the algorithm presented in Section 7. Whenever possible we solved step (7.2) using analytical formulas and resorted to the solver CVX (*http://cvxr.com/cvx/*) in the other cases. For example, in the case of the wedge penalty, we found that the computational time of the algorithm in Figure 4 is $495, 603, 665, 869, 1175$ faster than that of the solver CVX for $n = 100, 500, 1000, 2500, 5000$, respectively. Our implementation ran on a 16GM memory dual core Intel machine. The MATLAB code is available at http://www.cs.ucl.ac.uk/staff/M.Pontil/software.html.

Box. In the first experiment the model is 10-sparse, where each nonzero component, in a random position, is an integer uniformly sampled in the interval $[-10, 10]$. We wish to show that the more accurate the prior information about the model is, the more precise the estimate will be.



(a) (b) (c) (d)

Figure 5: Comparison between different penalty methods: (a) Box vs. Lasso; (b,c) Wedge vs. Hierarchical group Lasso; (d) Composite wedge. See text for more information

We use a box penalty (see Theorem 3.1) constructed "around" the model, imagining that an oracle tells us that each component $|\beta_i^*|$ is bounded within an interval. We consider three boxes $B[a,b]$ of different sizes, namely $a_i = (r - |\beta_i^*|)_+$ and $b_i = (|\beta_i^*| - r)_+$ and radii $r = 5, 1$ and 0.001, which we denote as Box-A, Box-B and Box-C, respectively. We compare these methods with the Lasso – see Figure 5-a. As expected, the three box penalties perform better. Moreover, as the radius of a box diminishes, the amount of information about the true model increases, and the performance improves.

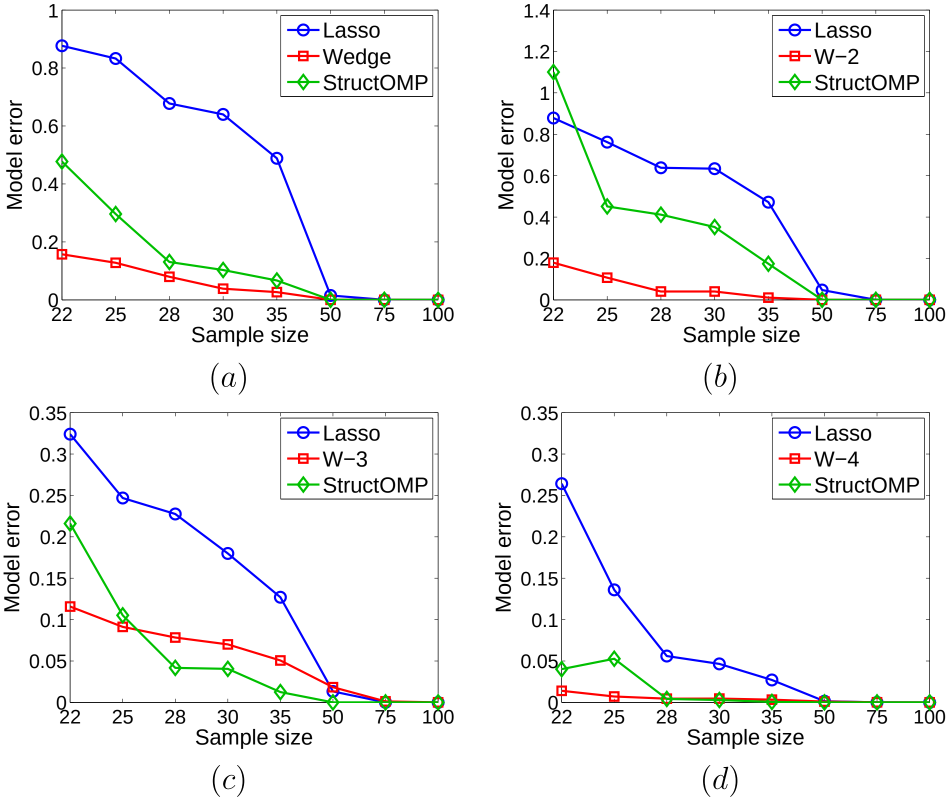
Wedge. In the second experiment, we consider a regression vector, whose components are nonincreasing in absolute value and only a few are nonzero. Specifically, we choose a 10-sparse vector: $\beta_j^* = 11 - j$, if $j \in \mathbb{N}_{10}$ and zero otherwise. We compare the Lasso, which makes no use of such ordering information, with the wedge penalty $\Omega(\beta|W)$ (see Theorem 4.1) and the hierarchical group Lasso in [32], which both make use of such information. For the group Lasso we choose $\Omega(\beta) = \sum_{\ell \in \mathbb{N}_{100}} \|\beta_{|J_\ell}\|_2$, with $J_\ell = \{\ell, \ell+1, \dots, 100\}$, $\ell \in \mathbb{N}_{100}$. These two methods are referred to as "Wedge" and "GL-lin" in Figure 5-b, respectively. As expected both methods improve over the Lasso, with "GL-lin" being the best of the two. We further tested the robustness of the methods, by adding two additional nonzero components with value of 10 to the vector β^* in a random position between 20 and 100. This result, reported in Figure 5-c, indicates that “GL-lin” is more sensitive to such a perturbation.

Composite wedge. Next we consider a more complex experiment, where the regression vector is sparse within different contiguous regions $P_1, \dots, P_{10}$, and the ℓ_1 norm on one region is larger than the ℓ_1 norm on the next region. We choose sets $P_i = \{10(i-1)+1, \dots, 10i\}$, $i \in \mathbb{N}_{10}$ and generate a 6-sparse vector β^* whose i-th nonzero element has value $31 - i$ (decreasing) and is in a random position in P_i, for $i \in \mathbb{N}_6$. We encode this prior knowledge by choosing $\Omega(\beta|\Lambda)$ with $\Lambda = \left\{\lambda \in \mathbb{R}^{100} : \|\lambda_{P_i}\|_1 \geq \|\lambda_{P_{i+1}}\|_1,\ i \in \mathbb{N}_9\right\}$. This method constraints the sum of the sets to be nonincreasing and may be interpreted as the composition of the wedge set with an average operation across the sets P_i, which may be computed using Proposition 2.3 . This method, which is referred to as “C-Wedge” in Figure 5-d, is compared to the Lasso and to three other versions of the group Lasso. The first is a standard group Lasso with the nonoverlapping groups $J_i = P_i$, $i \in \mathbb{N}_{10}$, thus encouraging the presence of sets of zero elements, which is useful because there are 4 such sets. The second is a variation of the hierarchical group Lasso discussed above with $J_i = \cup_{j=i}^{10} P_j$, $i \in \mathbb{N}_{10}$. A problem with these approaches is that the ℓ_2 norm is applied at the level of the individual sets P_i, which does not promote sparsity within these sets. To counter this effect we can enforce contiguous nonzero patterns within each of the P_i, as proposed by [14]. That is, we consider as the groups the sets formed by all sequences of $q \in \mathbb{N}_9$ consecutive elements at the beginning or at the end of each of the sets P_i, for a total of 180 groups. These three groupings will be referred to as “GL-ind”, “GL-hie’‘, “GL-con” in Figure 5-d, respectively. This result indicates the advantage of “C-Wedge” over the other methods considered. In particular, the group Lasso methods fall behind our method and the Lasso, with “GL-con” being slightly better than “GL-ind” and “GL-hie”. Notice also that all group Lasso methods gradually diminish the model error until they have a point for each dimension, while our method and the Lasso have a steeper descent, reaching zero at a number of points which is less than half the number of dimensions.

Polynomials. The constraints on the finite differences (see equation (4.10)) impose a structure on the sparsity of the model. To further investigate this possibility we now consider some models whose absolute value belong to the sets of constraints W^k, where $k = 1, \dots, 4$. Specifically, we evaluate the polynomials $p_1(t) = -(t+5)$, $p_2(t) = (t+6)(t-2)$, $p_3(t) = -(t+6.5)t(t-1.5)$ and $p_4(t) = (t+6.5)(t-2.5)(t+1)t$ at 100 equally spaced (0.1) points starting from -7. We take the positive part of each component and scale it to 10, so that the results can be seen in Figure 7. The roots of the polynomials has been chosen so that the sparsity of the models is either 18 or 19.

We solve the interpolation problem using our method with the penalty $\Omega(\beta|W^k)$, $k = 1, \dots, 4$, with the objective of testing the robustness of our method: the constraint set W^k should be a more meaningful choice when $|\beta^*|$ is in it, but the exact knowledge of the degree is not necessary. This is indeed the case: “W-k” outperforms the Lasso for every k, but among these methods the best one “knows” the degree of $|\beta^*|$. For clarity, in Figures 6 we included only the best method.

One important feature of these sparsity patterns is the number of contiguous regions: 1, 2, 2 and 3 respectively. This prior information cannot be exploited with convex optimization techniques, so we tested our method against StructOMP, proposed by [11], a state of the art greedy algorithm. It relies on a complexity parameter which depends on the number of contiguous re-



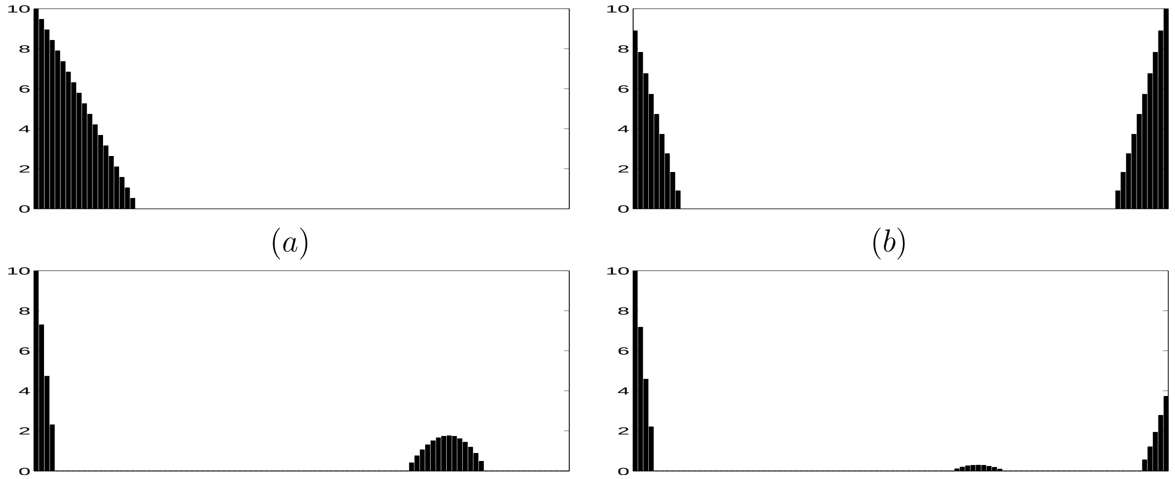
(a) (b) (c) (d)

Figure 6: Comparison between StructOMP and penalty $\Omega(\beta|W^k)$, $k = 1, \ldots, 4$, used for several polynomial models: (a) degree 1, (b) degree 2, (c) degree 3; (d) degree 4.

gions of the model, and which we provide exactly to the algorithm. The performance of “W-k” is comparable or better than StructOMP.

As a way of testing the methods on a less artificial setting, we repeat the experiment using the same sparsity patterns, but replacing each nonzero component with a uniformly sampled random number between 1 and 2. In Figure 8 we can see that, even if now the models manifestly do not belong to W^k, we still have an advantage because the constraints look for a limited number of contiguous regions. We found that in this case StructOMP has difficulties, probably due to the randomness of the model.

Finally, Figure 9 displays the regression vector found by the Lasso and the vector learned by “W-2” (left) and by the Lasso and “W-3” (right), in a single run with sample size of 15 and 35, respectively. The estimated vectors (green) are superposed to the true vector (black). Our method provides a better estimate than the Lasso in both cases. We found that the estimates of StructOMP are too variable for it to be meaningful to include one of them here.



(a) (b) (c) (d)

Figure 7: Silhouette of the polynomials by number of degree: (a) $k = 1$, (b) $k = 2$, (c) $k = 3$, (d) $k = 4$.

9 Conclusion

We proposed a family of penalty functions that can be used to model structured sparsity in linear regression. We provided theoretical, algorithmic and computational information about this new class of penalty functions. Our theoretical observations highlight the generality of this framework to model structured sparsity. An important feature of our approach is that it can deal with richer model structures than current approaches while maintaining convexity of the penalty function. Our practical experience indicates that these penalties perform well numerically, improving over state of the art penalty methods for structure sparsity, suggesting that our framework is promising for applications.

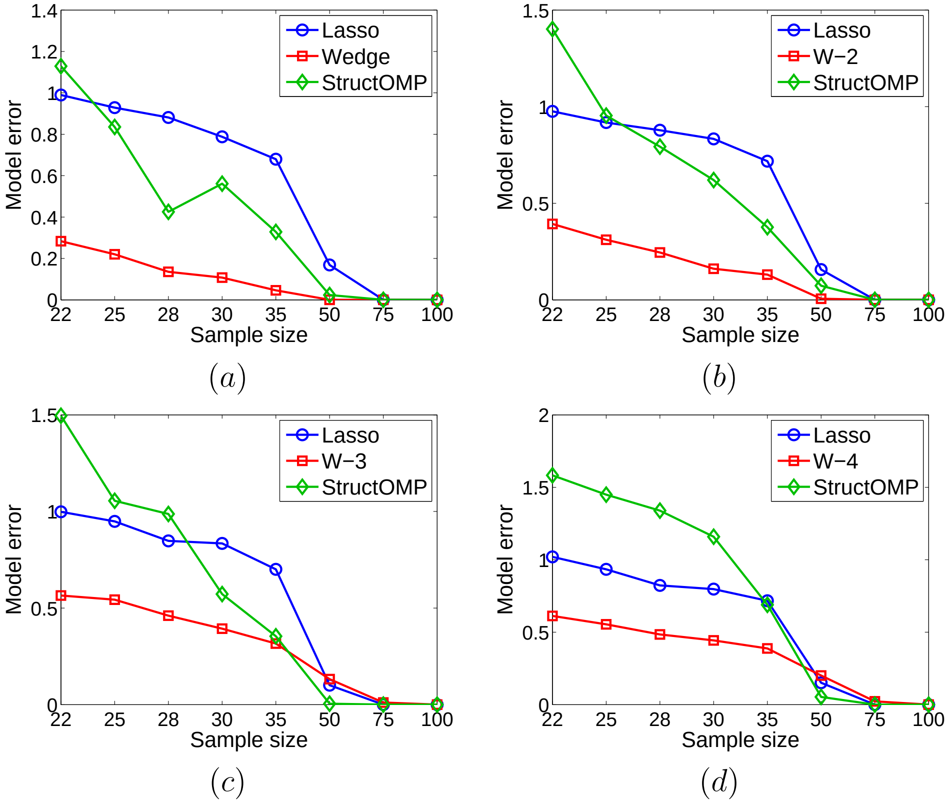
The methods developed here can be extended in different directions. We mention here several possibilities. For example, for any $r > 0$, it readily follows that

$$\|\beta\|_p^p = \inf\left\{ \frac{r}{r+1} \sum_{i \in \mathbb{N}_n} \frac{\beta_i^2}{\lambda_i} + \frac{1}{r}\lambda_i^r : \lambda \in \mathbb{R}_{++}^n \right\} \tag{9.1}$$

where $p = 2r/(r+1)$ and $\|\beta\|_p$ is the usual ℓ^p-norm on $\mathbb{R}^n$. This formula leads us to consider the same optimization problem over a constraint set Λ. Note that if $p \to 0$ the left hand side of the above equation converges to the cardinality of the support of the vector β.

Problems associated with multi-task learning [1, 2] demand matrix analogs of the results discussed here. In this regard, we propose the following family of unitarily invariant norms on $d \times n$ matrices. Let $k = \min(d, n)$ and $\sigma(B) \in \mathbb{R}_+^k$ be the vector formed from the singular values of B. When Λ is a nonempty convex set which is invariant under permutations our point of view in this paper suggests the penalty

$$\|B\|_\Lambda = \Omega(\sigma(B)|\Lambda).$$



(a) (b) (c) (d)

Figure 8: Comparison between StructOMP and penalty $\Omega(\beta|W^k)$, $k = 1, \ldots, 4$, used for several polynomial models with random values between the roots: (a) degree 1, (b) degree 2, (c) degree 3; (d) degree 4.

The fact that this is a norm, follows from the von Neumann characterization of unitarily invariant norms. When $\Lambda = \mathbb{R}^k_{++}$ this norm reduces to the trace norm [2].

Finally, the ideas discussed in this paper can be used in the context of kernel learning, see [3, 16, 17, 20, 25] and references therein. Let K_ℓ, $\ell \in \mathbb{N}_n$ be prescribed reproducing kernels on a set $\mathcal{X}$, and H_ℓ the corresponding reproducing kernel Hilbert spaces with norms $\|\cdot\|_\ell$. We consider the problem

$$\min\left\{\sum_{i\in\mathbb{N}_m}\left(y_i - \sum_{\ell\in\mathbb{N}_n} f_\ell(x_i)\right)^2 + \rho\Omega^2\Big((\|f_\ell\|_\ell : \ell \in \mathbb{N}_n)|\Lambda\Big) : f_\ell \in H_\ell, \ell \in \mathbb{N}_n\right\}$$

and note that the choice $\Lambda = \mathbb{R}^n_{++}$ corresponds to multiple kernel learning.

All the above examples deserve a detailed analysis and we hope to provide such in future work.

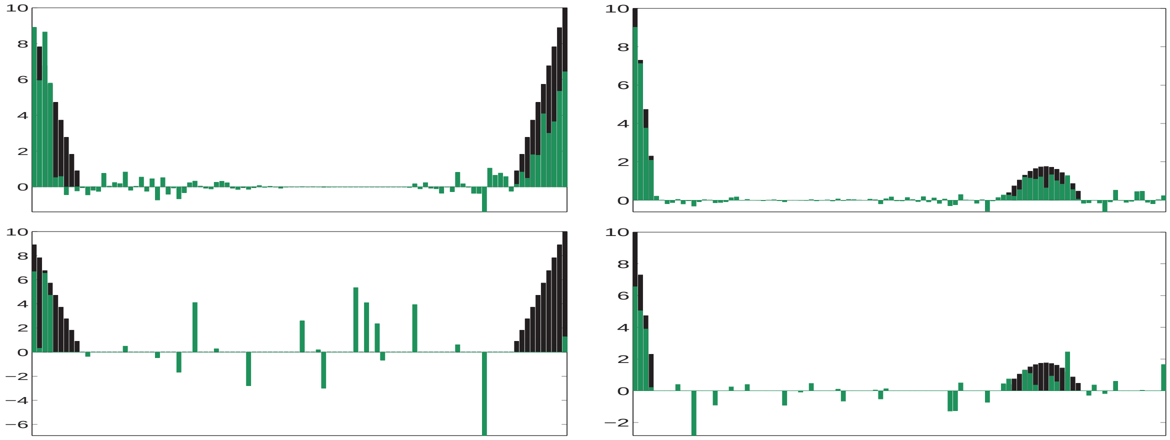


Figure 9: Lasso vs. penalty $\Omega(\cdot|\Lambda)$ for Convex (left) and Cubic (Right); see text for more information.

Acknowledgements

We are grateful to A. Argyriou for valuable discussions, especially concerning the proof of Theorem 7.1 and Theorem A.2 in the appendix. We also wish to thank Luca Baldassarre, Mark Herbster, Andreas Maurer, Raphael Hauser, Alexandre Tsybakov and Yiming Ying for useful discussions. This work was supported by Air Force Grant AFOSR-FA9550, EPSRC Grants EP/D071542/1 and EP/H027203/1, NSF Grant ITR-0312113, Royal Society International Joint Project Grant 2012/R2, as well as by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

A Appendix

In this appendix we describe in detail a result due to J.M. Danskin, which we use in the proof of Proposition 2.1.

Definition A.1. *Let f be a real-valued function defined on an open subset X of $\mathbb{R}^n$ and $u \in \mathbb{R}^n$. The directional derivative of f at $x \in X$ in the "direction" u is denoted by $(D_u f)(x)$ and is defined as*

$$(D_u f)(x) := \lim_{t \to 0} \frac{f(x+tu) - f(x)}{t}$$

if the limit exists. When the limit is taken through nonnegative values of t, we denote the corresponding right directional derivative by D_u^+.

Let Y be a compact metric space, $F : X \times Y \to \mathbb{R}$ a continuous function on its domain and define the function $f : X \to \mathbb{R}$ at $x \in X$ as

$$f(x) = \min \{F(x,y) : y \in Y\}.$$

We say that F is Danskin function if, for every $u \in \mathbb{R}^n$, the function $F'_u : X \times Y \to \mathbb{R}$ defined at $(x,y) \in X \times Y$ as $F'_u(x,y) = (D_u F(\cdot,y))(x)$ is continuous on $X \times Y$. Our notation is meant to convey the fact that the directional derivative is taken relative to the first variable of F.

Theorem A.1. *If X is an open subset of $\mathbb{R}^n$, Y a is compact metric space, $F : X \times Y$ is a Danskin function, $u \in \mathbb{R}^n$ and $x \in X$, then*

$$(D_u^+ f)(x) = \min \{F_u'(x,y) : y \in Y_x\}$$

where $Y_x := \{y : y \in Y,\ F(x,y) = f(x)\}$.

Proof. If $x \in X$, $y \in Y_x$ and $u \in \mathbb{R}^n$ then, for all positive t, sufficiently small, we have that

$$\frac{f(x+tu) - f(x)}{t} \leq \frac{F(x+tu,y) - F(x,y)}{t}.$$

Letting $t \to 0^+$, we get that

$$\limsup_{t \to 0^+} \frac{f(x+tu) - f(x)}{t} \leq \min \{F_u'(x,y) : y \in Y_x\}. \tag{A.1}$$

Next, we choose a sequence $\{t_k : k \in \mathbb{N}\}$ of positive numbers such that $\lim_{k\to\infty} t_k = 0$ and

$$\lim_{k \to \infty} \frac{f(x+t_k u) - f(x)}{t_k} = \liminf_{t \to 0^+} \frac{f(x+tu) - f(x)}{t}.$$

From the definition of the function f, there exists a $y_k \in Y$ such that $f(x + t_k u) = F(x + t_k u, y_k)$. Since Y is a compact metric space, there is a subsequence $\{y_{k_\ell} : \ell \in \mathbb{N}\}$ which converges to some $y_\infty \in Y$. It readily follows from our hypothesis that the function f is continuous on X. Indeed, we have, for every $x_1, x_2 \in X$, that

$$|f(x_1) - f(x_2)| \leq \max \{|F(x_1,y) - F(x_2,y)| : y \in Y\}.$$

Hence we conclude that $y_\infty \in Y_x$. Moreover, we have that

$$\frac{f(x+t_k u) - f(x)}{t_k} \geq \frac{F(x+t_k u, y_k) - F(x, y_k)}{t_k}.$$

By the mean value theorem, we conclude that there is positive number $\sigma_k < t_k$ such that the

$$\frac{f(x+t_k u) - f(x)}{t_k} \geq F_u'(x + \sigma_k u, y_k).$$

We let $\ell \to \infty$ and use the hypothesis that F is a Danskin function to conclude that

$$\liminf_{t \to 0^+} \frac{f(x+tu) - f(x)}{t} \geq F_u'(x, y_\infty) \geq \min \{F_u'(x,y) : y \in Y_x\}.$$

Combining this inequality with (A.1) proves the result. ■

We note that [4, p. 737] describes a result which is attributed to Danskin without reference. That result differs from the result presented above. The result in [4, p. 737] requires the hypothesis of convexity on the function F. The theorem above and its proof is an adaptation of Theorem 1 in [8].

We are now ready to present the proof of Proposition 2.1.

Proof of Proposition 2.1 The essential part of the proof is an application of Theorem A.1. To apply this result, we start with a $\beta \in (\mathbb{R}\backslash\{0\})^n$ and introduce a neighborhood of this vector defined as

$$X(\beta) = \left\{\alpha : \alpha \in \Lambda, \|\alpha - \beta\|_\infty < \frac{\beta_{\min}}{2}\right\},$$

where $\beta_{\min} = \min\{|\beta_i| : i \in \mathbb{N}_n\}$. Theorem A.1 also requires us to specify a compact subset $Y(\beta)$ of $\mathbb{R}^n$. We construct this set in the following way. We choose a fixed $\overline{\lambda} \in \Lambda$ and a positive $\epsilon > 0$. From these constants we define the constants

$$\begin{aligned} c(\beta) &= \sum_{i \in \mathbb{N}_n} \left(\frac{(|\beta_i| + \beta_{\min}/2)^2}{\overline{\lambda}_i} + \overline{\lambda}_i \right), \\ a(\beta) &= \frac{\beta_{\min}^2}{4(c(\beta) + \epsilon)}, \\ b(\beta) &= \max(a(\beta), c(\beta) + \epsilon). \end{aligned}$$

With these definitions, we choose our compact set $Y(\beta)$ to be $Y(\beta) = \Lambda_{a(\beta),b(\beta)}$. To apply Theorem A.1, we use the fact, for any $\alpha \in X(\beta)$, that

$$\Omega(\alpha|\Lambda) = \min\{\Gamma(\alpha, \lambda) : \lambda \in Y(\beta)\}. \tag{A.2}$$

Let us, for the moment, assume the validity of this equation and proceed with the remaining details of the proof. As a consequence of this equation, we conclude that there exists a vector $\lambda(\beta)$ such that $\Omega(\beta|\Lambda) = \Gamma(\beta, \lambda(\beta))$. Moreover, when $\beta \in (\mathbb{R}\backslash\{0\})^n$ the function $\Gamma_\beta : \mathbb{R}^n_{++} \to \mathbb{R}$, defined for $\lambda \in \mathbb{R}^n_{++}$, as $\Gamma_\beta(\lambda) = \Gamma(\beta, \lambda)$ is strictly convex on its domain and so, $\lambda(\beta)$ is unique.

By construction, we know, for every $\alpha \in X(\beta)$, that

$$\max\left\{\left|\lambda_i(\alpha) - \frac{a(\beta) + b(\beta)}{2}\right| : i \in \mathbb{N}_n\right\} \leq \frac{a(\beta) + b(\beta)}{2}.$$

From this inequality we shall establish that $\lambda(\beta)$ depends continuously on β. To this end, we choose any sequence $\{\beta^k : k \in \mathbb{N}\}$ which converges to β and from the above inequality we conclude that the sequence of vectors $\lambda(\beta^k)$ is bounded. However this sequence can only have one cluster point, namely $\lambda(\beta)$, because Γ is continuous. Specifically, if $\lim_{k\to\infty} \lambda(\beta^k) = \tilde{\lambda}$, then, for every $\lambda \in \Lambda$, it holds that $\Gamma(\beta^k, \lambda(\beta^k)) \leq \Gamma(\beta^k, \lambda)$ and, passing to the limit $\Gamma(\beta, \tilde{\lambda}) \leq \Gamma(\beta, \lambda)$, implying that $\tilde{\lambda} = \lambda(\beta)$.

Likewise, equation (A.2) yields the formula for the partial derivatives of $\Omega(\cdot|\Lambda)$. Specifically, we identify F and f in Theorem A.1 with Γ and $\Omega(\cdot|\Lambda)$, respectively, and note that

$$\frac{\partial\Omega}{\partial\beta_i}(\beta|\Lambda) = \min\left\{\frac{\partial\Gamma}{\partial\beta_i}(\beta, \lambda) : \lambda \in \Lambda,\ \Gamma(\beta, \lambda) = \Omega(\beta|\Lambda)\right\} = \frac{\partial\Gamma}{\partial\beta_i}(\beta, \lambda(\beta)) = 2\frac{\beta_i}{\lambda_i(\beta)}.$$

Therefore, the proof will be completed after we have established equation (A.2). To this end, we note that if $\lambda = (\lambda_i : i \in \mathbb{N}_n) \in \Lambda \backslash Y(\beta)$ then there exists $j \in \mathbb{N}_n$ such that either $\lambda_j < a(\beta)$ or $\lambda_j > b(\beta)$. Thus, we have, for every $\alpha \in X(\beta)$, that

$$\Gamma(\alpha, \lambda) \geq \frac{1}{2}\left(\frac{\alpha_j^2}{\lambda_j} + \lambda_j\right) \geq \frac{1}{2}\min\left(\frac{\beta_{\min}^2}{4a(\beta)}, b(\beta)\right) = \frac{c(\beta)+\epsilon}{2} \geq \Omega(\alpha|\Lambda) + \frac{\epsilon}{2}.$$

This inequality yields equation (A.2). ■

We end this appendix by extracting the essential features of the convergence of the alternating algorithm as described in Section 7. We start with two compact sets, $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$, and a strictly convex function $F : X \times Y \to \mathbb{R}$. Corresponding to F we introduce two additional functions, $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ defined, for every $x \in X, y \in Y$ as

$$f(x) = \min\{F(x, y') : y' \in Y\}, \quad g(y) = \min\{F(x', y) : x' \in X\}.$$

Moreover, we introduce the mappings $\phi_1 : Y \to X$ and $\phi_2 : X \to Y$, defined, for every $x \in X$, $y \in Y$, as

$$\phi_1(y) = \text{argmin}\{F(x,y) : x \in X\}, \quad \phi_2(x) = \text{argmin}\{F(x,y) : y \in Y\}.$$

Lemma A.1. *The mappings ϕ_1 and ϕ_2 are continuous on their respective domain.*

Proof. We prove that ϕ_1 is continuous. The same argument applies to ϕ_2. Suppose that $\{y^k : k \in \mathbb{N}\}$ is a sequence in Y which converges to some point $y \in Y$. Then, since F is jointly strictly convex, the sequence $\{\phi_1(y^k) : k \in \mathbb{N}\}$ has only one cluster point in X, namely $\phi_1(y)$. Indeed, if there is a subsequence $\{\phi_1(y^{k_\ell}); \ell \in \mathbb{N}\}$ which converges to $\tilde{x}$, then by definition, we have, for every $x \in X$, $\ell \in \mathbb{N}$, that $F(\phi_1(y^{k_\ell}), y^{k_\ell}) \leq F(x, y^{k_\ell})$. From this inequality it follows that $F(\tilde{x}, y) \leq F(x, y)$. Consequently, we conclude that $\tilde{x} = \phi_1(y)$. Finally, since X is compact, we conclude that the $\lim_{k\to\infty} \phi_1(y^k) = \phi_1(y)$. ■

As an immediate consequence of the lemma, we see that f and g are continuous on their respective domains, because, for every $x \in X, y \in Y$, we have that $f(x) = F(x, \phi_2(x))$ and $g(y) = F(\phi_1(y), y)$.

We are now ready to define the alternating algorithm.

Definition A.2. *Choose any $y_0 \in \text{int}(Y)$ and, for every $k \in \mathbb{N}$, define the iterates*

$$x^k = \phi_1(y^{k-1})$$

and

$$y^k = \phi_2(x^k).$$

Theorem A.2. *If $F : X \times Y \to \mathbb{R}$ satisfies the above hypotheses and it is differentiable on the interior of its domain, and there are compact subsets $X_0 \subset \text{int}(X)$, $Y_0 \subseteq \text{int}(Y)$ such that, for all $k \in \mathbb{N}$, $(x^k, y^k) \in X_0 \times Y_0$, then the sequence $\{(x^k, y^k) : k \in \mathbb{N}\}$ converges to the unique minimum of F on its domain.*

Proof. First, we define, for every $k \in \mathbb{N}$, the real numbers $\theta_k = F(x^k, y^{k-1})$ and $\nu_k = F(x^k, y^k)$. We observe, for all $k \geq 2$, that

$$\nu_k \leq \theta_k \leq \nu_{k-1}.$$

Therefore, there exists a constant ψ such that $\lim_{k\to\infty} \theta_k = \lim_{k\to\infty} \nu_k = \psi$. Suppose, there is a subsequence $\{x^{k_\ell} : \ell \in \mathbb{N}\}$ such that $\lim_{\ell\to\infty} x^{k_\ell} = x$. Then $\lim_{\ell\to\infty} \phi_2(x^{k_\ell}) = \phi_2(x) =: y$. Observe that $\nu_k = f(x^k)$ and $\theta_{k+1} = g(y^k)$. Hence we conclude that

$$f(x) = g(y) = \psi.$$

Since F is differentiable, (x, y) is a stationary point of F in $\text{int}(X) \times \text{int}(Y)$. Moreover, since F is strictly convex, it has a unique stationary point which occurs at its global minimum. ■

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] A. Argyriou, C.A. Micchelli, and M. Pontil. On spectral learning. *The Journal of Machine Learning Research*, 11:935–953, 2010.

[3] F. R. Bach, G. R. G Lanckriet, and M. I. Jordan. Multiple kernels learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[4] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[5] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[8] J.M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

[9] A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *IEEE International Symposium on Biomedical Imaging*, 2009.

[10] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. 2009.

[11] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 417–424. ACM, 2009.

[12] L. Jacob. *Structured Priors for Supervised Learning in Computational Biology*. 2009. Ph.D. Thesis.

[13] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *International Conference on Machine Learning (ICML 26)*, 2009.

[14] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. arXiv:0904.3523v2, 2009.

[15] S. Kim and E.P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 543–550. Omnipress, 2010.

[16] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38(6):3660–3695, 2010.

[17] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[18] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

[19] K. Lounici, M. Pontil, A.B. Tsybakov, and S. Van De Geer. Oracle Inequalities and Optimal Inference under Group Sparsity. *Arxiv preprint arXiv:1007.1771*, 2010.

[20] C. A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66:297–319, 2007.

[21] C.A. Micchelli, J.M. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1612–1623. 2010.

[22] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving Structured Sparsity Regularization with Proximal Methods. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010)*, pages 418–433, 2010.

[23] G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):1–22, 2010.

[24] A.B. Owen. A robust hybrid of lasso and ridge regression. In *Prediction and discovery: AMS-IMS-SIAM Joint Summer Research Conference, Machine and Statistical Learning: Prediction and Discovery*, volume 443, page 59, 2007.

[25] T. Suzuki R. Tomioka. Regularization strategies and empirical bayesian learning for MKL. arXiv:1001.26151, 2011.

[26] F. Rapaport, E. Barillot, and J.P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, 2008.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[28] S.A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614, 2008.

[29] Z. Xiang, Y. Xi, U. Hasson, and P. Ramadge. Boosting with spatial regularization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2107–2115. 2009.

[30] M. Yuan, R. Joseph, and H. Zou. Structured variable selection and estimation. *Annals of Applied Statistics*, 3(4):1738–1757, 2009.

[31] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[32] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.